KEI
Working
Paper
2017-05

환경 분야 빅데이터 수집방법 연구

대기질 데이터를 중심으로

한국진 · 강성원 · 김도연 · 김영인

KEI 한국환경정책·평가연구원
Korea Environment Institute

연구진

연구책임자 한국진 (한국환경정책·평가연구원 선임전문원)

참여연구원 강성원 (한국환경정책·평가연구원 선임연구위원)

김도연 (한국환경정책·평가연구원 연구원)

김영인 (한국환경정책·평가연구원 선임전문원)

연구자문위원 (가나다 순)

강희찬 (인천대학교 교수)

김종호 (한국환경정책·평가연구원 선임연구위원)

명수정 (한국환경정책·평가연구원 연구위원)

이명진 (한국환경정책·평가연구원 부연구위원)

발행인 이 창 훈 (원장 직무대리)

발행처 한국환경정책·평가연구원
(30147) 세종특별자치시 시청대로 370
세종국책연구단지 과학·인프라동
전화 044-415-7777 팩스 044-415-7799
http://www.kei.re.kr

인 쇄 2017년 9월 26일

발 행 2017년 9월 30일

등 록 제 2015-000009호 (1998년 1월 30일)

ISBN 979-11-5980-127-3 93530

이 보고서를 인용 및 활용 시 아래와 같이 출처를 표시해 주십시오.
한국진 외(2017), 「환경 분야 빅데이터 수집방법 연구: 대기질 데이터를 중심으로」, 한국환경정책·평가연구원.

값 5,000원

서 언

지능정보화를 통한 환경오염 문제 해결, 친환경 이동수단 및 기반 마련, 신재생·스마트 에너지 인프라 구축, 미세먼지 걱정 없는 쾌적한 대기환경 조성, 지속가능한 국토 환경 조성, 환경친화형 농수산업으로의 전환 등 올해 정부 시책만 봐도 환경은 다양한 분야에서 주요 이슈이자 핵심 가치입니다.

올바른 환경정책을 입안하기 위해서는 시의성 있고 적절한 환경 연구가 수행되어야 합니다. 이에 연구에 활용할 수 있는 환경 분야 빅데이터 수집방법을 제시한 본 연구는 시의적절하다고 볼 수 있습니다. 또한 환경 빅데이터 수집-저장 절차부터 프레임워크(안)까지 제시하여 다양한 환경 빅데이터 연구에 적용할 수 있어 향후 환경 연구 수행에 중요한 바탕이 될 것으로 기대됩니다.

끝으로 본 연구를 수행한 한국환경정책·평가연구원 미래환경연구본부 빅데이터연구팀의 한국진 선임전문원, 강성원 박사, 김도연 연구원, 김영인 선임전문원께 감사를 표합니다. 바쁘신 와중에도 자문을 통해 연구에 도움을 주신 인천대학교 강희찬 교수께 깊은 감사를 드립니다. 또한 우리 원의 김종호 박사, 명수정 박사, 이명진 박사의 자문에도 감사의 마음을 전합니다.

2017년 9월

한국환경정책·평가연구원

원장 직무대리 **이 창 훈**

국문요약

본 연구는 지능정보사회의 근간인 빅데이터에 대한 이해를 통해 환경 연구에 활용 가능한 빅데이터를 식별하고 데이터 기반 연구혁신을 위한 수집 방법으로서 환경 빅데이터 수집-저장의 절차와 프레임워크(안)를 제시하였다. 미래 사회와 연구 패러다임의 중심에 선 빅데이터를 환경 연구에 활용하기 위해서는 빅데이터에 대한 충분한 이해와 적극적인 활용이 필요하다. 더불어 환경 분야 빅데이터에 대한 식별 및 대응(안)도 마련되어야 한다. 이에 대한 사례로서 한국환경공단의 대기질 빅데이터 및 그 서비스를 분석하였고 분석과정을 통해 빅데이터 수집-저장방법의 절차를 검토하고 수집방법에 대한 프레임워크(안)를 제시하였다. 본 연구의 주요 내용은 다음의 3가지로 요약할 수 있다.

○ 빅데이터의 이해

빅데이터는 데이터의 수집-저장-분석-(시각화)-예측의 절차를 갖고 있지만 사회 전반적으로 다양한 이해와 의미를 갖고 있어 환경 분야 빅데이터 또한 다른 접근방법 및 이해가 필요하다. 그동안 빅데이터가 정부 주도형으로 추진되어 양적 성장을 이뤄냈고 우리나라에서는 공공데이터포털을 통해 데이터가 없더라도 데이터 제공자를 찾을 수 있는 제도적 장치가 마련되어 있다. 그러나 데이터 처리를 위한 첫 번째 단계인 수집-저장 단계에서는 성장보다 접근성, 활용성이 요구되며 연구자의 애로사항이나 수요를 고려한 데이터를 활용할 수 있는 수요자 관점에서 수집방법을 검토하였다.

○ 환경 분야 빅데이터

환경 분야의 빅데이터라 함은 모든 분야의 데이터를 일컫는다고 해도 과언이 아니다. 따라서 수요자 중심의 데이터 우선순위를 부여하고 그 사례를 제시하였다. 공공데이터포털 활용 신청 순위의 검토 및 한국환경정책·평가연구원 연구자 대상의 데이터 활용 온라인 설문을 통해 기상기후 및 대기질 데이터가 도출되었다. 이 가운데 활용성이 우수하고 동일한 규모의 데이터셋을 제공하고 있는 한국환경공단의 대기질 데이터 및 데이터 서비스를 분석하였다.

○ 수집방법의 절차화

위와 같은 분석을 통해 연구자들에게 특정 빅데이터에 대한 수집방법만을 제시한다면 기존의 방법과 다르지 않다고 판단하였다. 이에 도출된 수집방법을 통해 수집-저장방법의 절차를 마련하고 이를 프레임워크(안)으로 제시하고자 하였다. 이를 활용하면 다른 환경 빅데이터를 활용하는 연구에도 적용할 수 있고 컴퓨팅 플랫폼에도 적용이 수월하다. 또한 빅데이터 수집-저장 프레임워크(안)를 통해 활용 가능한 구체적인 소프트웨어 등의 컴퓨팅 환경을 언급하여 데이터 기반 연구수행 체계로의 전환 또는 접근이 용이하도록 안을 제시하였다.

주제어 : 빅데이터, 수집방법, 저장방법, 웹크롤링, 프레임워크

차례

표차례

| 그림차례 |

| 약어 |

ML	Machine Learning (기계학습)
DL	Deep Learning (심층학습)

제1장

서 론

1. 연구의 필요성 및 목적

우리나라 환경 분야를 대표하는 정부 부처인 환경부와 환경정책 연구 및 환경영향평가서 검토를 수행하는 연구기관인 한국환경정책·평가연구원의 홈페이지[1] 검색창에서 '빅데이터'를 검색해보면 2014년부터 검색결과가 나타난다. 반면 가트너(Gartner, Inc.)가 매년 발표하는 10대 전략기술(Top 10 Strategic Technology Trends)에서는 2012년 빅데이터(Big Data), 2013년 전략적인 빅데이터(Strategic Big Data)만 등장했다가 바로 사라졌다.[2] 그럼에도 불구하고 최근 3년간 우리나라에서는 빅데이터가 환경 연구 분야뿐만 아니라 전 산업계, 학계 등 사회 전반적으로 이슈가 되고 있다. 최근 환경 분야의 연구보고서 제목을 보면 'DB' 보다 '데이터'가 많이 등장하고 '구축'보다 '활용'이나 '연계'가 자주 등장하는 추세이다. 또한 '데이터', '정보', '지식', 'DB'의 경계가 모호하지만 사용에 있어서는 미묘한 차이가 느껴진다. 동일한 의미의 '데이터'라 할지라도 목적이 아닌 수단으로서, 활용의 개념으로 빠르게 변화하고 있다. 그러나 환경 분야의 빅데이터는 넓은 분야를 다루는 그 영역의 특성 때문인지 다소 출발이 지연된 것처럼 느껴진다.

더불어 미래환경대응 정보화 전략 TF 결과보고서(2016)에서는 2015년까지의 한국환경정책·평가연구원이 수행한 연구과제의 약 24%(사업비 기준)가 정보화사업 성격으로 추진되었음에도 불구하고 연구보고서 및 연구 산출물을 공유·활용할 수 있는 데이터 아카이브[3]와

---

1) 환경부, http://www.me.go.kr; 한국환경정책평가연구원, http://www.kei.re.kr.

2) 가트너(Gartner, Inc.)가 발표하는 「10대 전략기술(Top 10 Strategic Technology Trends」은 기업들이 비즈니스 활동을 하는 데 중요한 영향을 미칠 것으로 예상되는 기술들을 예측한 보고서임.

3) 연구데이터 아카이브는 연구과정에서 생산되는 모든 데이터를 저장하고 해당 시점을 재생할 수 있는 시스템임.

연구과정 데이터들의 메타데이터 부재를 지적하였다. 정보화사업 성격으로 추진되는 연구과제가 다른 연구과제에 비해 갖는 이점은 선행 사례조사 및 데이터 재사용 등을 목적으로 연구 수행 과정의 검토(또는 검증) 및 재생이 수월하고 연구 수행 결과를 발전시키기 유리하다는 점이다. 정보화사업의 특성상 수많은 정보화사업 산출물들이 연구결과물을 뒷받침하고 이러한 산출물들은 연구과정을 복원하고 검증하는 데 도움이 되기 때문이다. 그러나 이러한 이점에도 불구하고 목적에 따라 또는 필요에 따라 삭제하거나 동일 데이터를 사용하는 연구 수행 시 연구에 활용할 데이터를 검색-수집-저장-전처리를 거쳐 연구에 활용하는 일이 반복되고 있는 것이다. 물론 완전히 동일한 데이터를 사용한다는 의미는 아니다. 그러나 많은 연구자들이 데이터를 찾고 저장하는 과정 또는 이후 다른 연구자가 활용할 수 있도록 데이터 저장소 및 데이터 공유체계가 있었으면 하는 바람을 갖고 있다. 이것은 선행 연구에 대한 데이터와 연구과정을 검토하고 재생할 수 있는 고품질 연구 데이터 아카이브와 환경 분야의 특성상 다양하고 방대한 빅데이터를 수집-저장할 수 있는 수집 방법 및 수집-저장 프레임워크가 요구됨을 의미한다.

그리고 환경부, 기상청, 국립환경공단, 국립환경과학원 등 환경 관련 빅데이터 생산기관에서는 파일, 데이터베이스(DBMS)[4], API[5] 형태의 데이터를 양이나 시간의 제약 없이 빅데이터 서비스를 통해 제공하고 있다. 그런데, 이를 이용하고자 하는 연구자는 자신이 활용하고자하는 데이터를 어느 기관에서 제공하는지, 어떤 서비스를 통해 해당 데이터를 사용할 수 있는지, 해당 데이터가 어떤 형태로 제공하는지 등 데이터를 식별하는 데 많은 노력이 든다. 막상 데이터를 찾았다고 해도 연구목적이나 연구자 성향, 선호하는 데이터 유형에 따라 데이터를 수집-저장하는 데 많은 어려움을 겪고 있다. 예를 들어, 국립환경공단이 제공하는 대기질 빅데이터는 공공데이터포털과 에어코리아를 통해 제공된다. 이것을 활용하고자 하는 연구자는 환경 연구 분야의 특성상 요구되는 시간해상도와 실시간 계측 및 확정 데이터를 구분하여 확보하고자 할 것이다. 현재 시점을 기준으로 과거 데이터를 수집-저장하는 경우에는 엑셀 등 파일이나 제이슨(이하 'JSON')[6]이나 XML[7] 형태의 오픈API 연동

4) 데이터베이스(DBMS)는 구조적으로 디지털화된 온라인 데이터를 의미하며 DB라고 총칭하기도 함.
5) API(Application Programming Interface)는 정보서비스 간 데이터 통로를 의미하며 개방형 API를 보통 오픈API(OpenAPI)라 부르며, API라 표시하면 개방형 API와 폐쇄형 API를 모두 이름.
6) JSON(JavaScript Object Notation)은 HTML(각주 13 참조)이나 XML(각주 7 참조)보다 간결하고 구조적인

데이터를 수집-저장할 것이고 현재 시점을 기준으로 미래 데이터(해당 시점의 실시간 데이터)를 수집-저장하는 경우에는 JSON이나 XML 형태의 오픈API 연동 데이터를 수집-저장할 것이다. 과거 데이터의 경우에는 다수의 연구자들이 중복된 데이터를 수집-저장하는 경우가 많고 R[8], 파이선(Python)[9], 자바(Java)[10] 등 프로그래밍 언어를 잘 모르거나 사용하는 소프트웨어가 오픈API 연동 데이터를 처리할 수 없는 연구자들은 다른 연구자의 도움을 필요로 한다. 따라서 환경 분야 빅데이터의 수집-저장 과정에 대한 중복수행을 방지하고 자동화할 수 있는 수집방법과 절차를 마련하고 연구 데이터 공유를 위해 수집-저장 방법과 절차를 체계화해야 한다.

환경 연구에 적합한 빅데이터 수집방법과 절차를 체계화하여 데이터를 수집-저장하면 연구자는 수집-저장 절차보다 본연의 수집 대상 데이터 분석 및 예측 작업에 집중할 수 있다. 또한 이를 프레임워크[11]로 만들면 컴퓨팅 플랫폼[12]에 반영하기 쉽고 데이터 품질을 유지할 수 있으며 연구 활용을 위한 빅데이터의 저장 및 적재도 수월해진다. 예를 들어, 연구자가 필요한 데이터를 제공하는 웹서비스에 접속해 검색조건을 설정하여 필요한 데이터를 조회하여 다운로드하는 경우, 웹서비스의 특성상 HTML[13]과 같은 웹서비스 구조를 잘 모르면 수십 번에서 수백 번 이상 클릭을 해서 데이터를 다운로드해야 하고, 또 이를 모아서 편집·저장하게 된다. 이와 같은 절차는 적게는 수 시간 내지 며칠에서 몇 주까지도 걸린다. 경우에 따라서 웹서비스 타임아웃(timeout)[14]이나 자동 로그아웃(logout) 등 예외

데이터 형태임. 데이터를 키와 값으로 단순화하여 대량전송 등 정보서비스 연동 시 많이 사용함.

7) XML(eXtensible Markup Language)은 사용목적에 따라 메타데이터를 기술할 수 있는 구조화된 문서 형태 및 논리적인 구조를 명시한 언어를 말함.

8) 1993년 오클랜드 대학교에서 개발된 통계 및 그래프 작업을 위한 오픈소스 인터프리터 프로그래밍 언어로 다양한 라이브러리와 쉬운 문법으로 다양한 분야에서 활용함.

9) 1989년 귀도 판 로썸(Guido van Rossum)이 창시하고 1991년 발표된 기존 프로그래밍 언어에 비해 쉬워 실 사용률과 생산성이 좋은 프로그래밍 언어, https://www.python.org.

10) 썬 마이크로시스템즈(Sun Microsystems)의 제임스 고슬링(James Gosling)이 개발하고 1995년 썬(2010년 오라클이 인수)이 발표한 플랫폼 독립적인 객체지향 프로그래밍 언어, https://go.java.

11) 프레임워크(Framework)는 연구과정을 재사용할 수 있도록 구체화하는 것을 의미함.

12) 컴퓨팅 플랫폼(Computing Platform)은 목적을 가진 소프트웨어를 이용할 수 있는 하드웨어와 소프트웨어를 의미함.

13) HTML(Hyper Text Mark-up Language)는 웹페이지의 형태 및 논리적인 구조를 명시한 언어를 말함

14) 웹서비스 타임아웃(timeout)이란 이용자가 일정 시간 요청이 없는 경우, 이용자가 설정한 조건이나 접속을 만료시키는 것을 말함.

처리 대응을 위해 자리를 비울 수 없는 경우가 많아 연구 활동을 위한 준비 작업에만 많은 시간과 인력 등 노력이 수반된다. 따라서 연구자가 연구 목적에 집중할 수 있도록 데이터 수집과정 및 절차가 반영된 수집-저장 프레임워크 구축(안)까지 검토가 필요하다.

2. 연구 범위

환경 연구에 빅데이터를 활용하기 위한 첫 번째 단계는 연구 목적에 부합하는 데이터를 결정하는 일이지만, 이 부분은 연구를 설계하는 연구자의 역량과 필요에 의해 결정되는 것이므로 빅데이터 수집-저장 프레임워크에 데이터 공유방법으로만 언급하였다. 두 번째는 선정된 데이터의 제공자를 찾아서 데이터를 수집-저장-(전처리)하여 연구에 활용 가능한 데이터로 재구축해야 한다. 이 경우 원본 데이터를 보존하고 수집한 데이터에 대한 정보와 수집방법, 즉 가공한 데이터에 대한 정보와 그 처리방법 등을 포함한 메타데이터를 생산하고 공유해야만 한다.

더불어 데이터 제공자나 데이터를 활용하는 연구자가 필요한 데이터에 쉽게 접근하기 위해서는 데이터의 분류체계가 필요하다. 환경 분야가 다양한 분야에 걸쳐있는 특성 때문에 환경 분야를 명확히 하고 이를 기반으로 일관된 데이터 분류 기준을 마련해야 한다. 본 연구에서는 모든 환경 분야 빅데이터에 대한 수집방법을 검토하기보다 연구 범위를 축소하는 것이 좋겠다는 자문의견을 받아들여 수요분석을 통해 대상 빅데이터를 검토하였다. 환경 분야의 범주가 물, 대기, 토양 등 자연환경으로 시작해 토목, 건축, 교통 등 인공환경까지 너무나 범위가 넓고 환경에만 해당되지 않는 세부 분야도 많다. 예를 들어 제10차 표준산업분류(KSIC)[15]를 보면 「E 수도, 하수 및 폐기물 처리, 원료 재생업」(세분류 36~39) 외 「F 건설업」 내 환경설비 건설업, 「M 천문, 과학 및 기술 서비스업」 내 72122 환경 관련 엔지니어링, 「O 공공 행정, 국방 및 사회보장 행정」 내 84213 환경 행정 서비스업으로 환경 분야를 분류하고 있다.

따라서 본 연구에서는 환경 분야의 빅데이터 분류체계를 정의하기보다 수요자 측면에서

15) 통계청 통계분류포털(https://kssc.kostat.go.kr)에서 제공하며 제10차 표준산업분류(KSIC)는 2017년 7월 1일부터 시행.

중요하다고 판단되는 빅데이터를 선정하고 그 데이터를 수집-저장하여 수집방법의 사례로 제시하였다. 또한 온라인 서비스가 제공되지 않거나 이미지 파일이나 PDF, HWP와 같이 다운로드 후 전처리 차수가 높은 데이터는 배제하였다. 이에 다음의 내용을 포함한 환경 분야 빅데이터 수집방법을 제시하였다.

- 웹페이지 및 RESTful 데이터[16]를 수집하기 위한 빅데이터 수집방법
- 데이터를 오픈소스 DBMS[17]에 저장 및 소프트웨어에 적재할 수 있는 수집방법
- 재사용 가능한 DB 스키마(구조)를 반영한 ERD[18] 제시

3. 연구 내용 및 수행 체계

본 연구에서는 환경 분야 빅데이터 중 수요(활용) 데이터 우선순위를 반영한 데이터를 기준으로 데이터를 수집-저장하는 수집방법과 그 절차를 프레임워크 구축(안)으로 제시하고자 한다. 이에 환경 분야 빅데이터 중 공공데이터포털과 한국환경정책·평가연구원에서의 수요(활용) 데이터 우선순위를 반영하여 한국환경공단의 대기질 데이터를 예시로 환경 분야 빅데이터를 수집-저장하는 수집방법을 제시하였다.

주요 연구 내용과 수행 체계는 다음과 같다(그림 1-1 참조).

첫째, 빅데이터 출현 배경 등의 흐름과 판단기준 및 국내 빅데이터 서비스 현황과 유형을 검토하여 환경 분야 빅데이터에 대해서 살펴보았다.

둘째, 국내 최대 빅데이터 서비스인 공공데이터포털의 데이터 유형과 이용방법에 대해 살펴보고 공공데이터포털의 빅데이터 수요와 한국환경정책·평가연구원의 데이터 수요를 검토하였다.

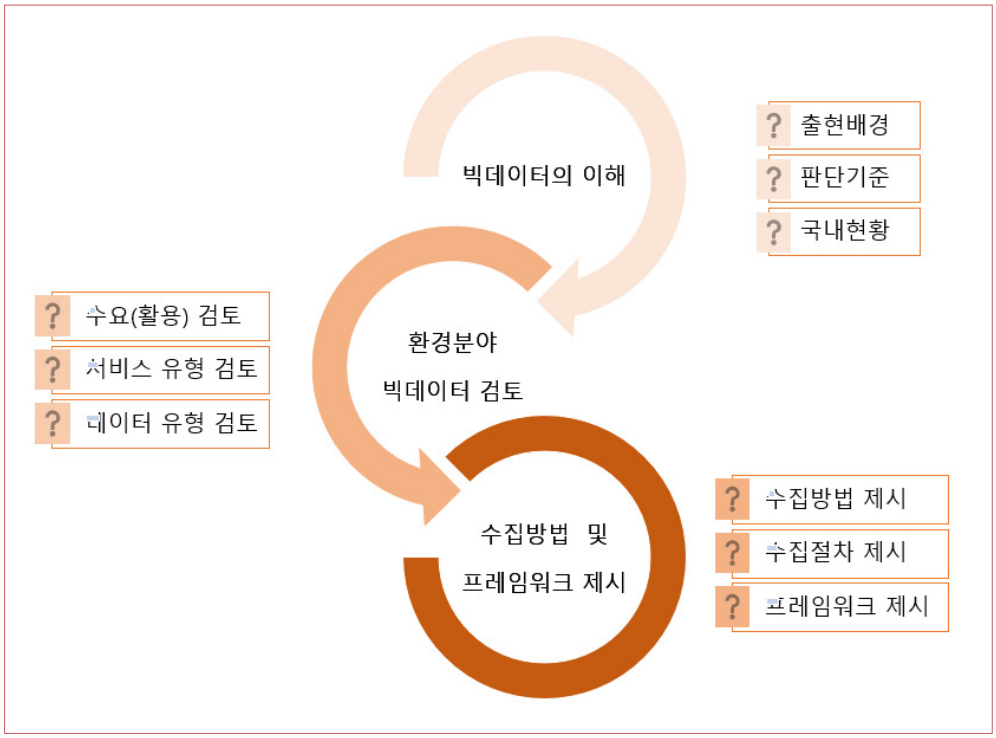
셋째, 수요(활용) 데이터 우선순위를 반영한 한국환경공단의 대기질 빅데이터 서비스와 제공되는 데이터 유형을 검토하여 해당 데이터를 수집-저장하는 수집방법을 도출하여 그 절차를 프레임워크(안)으로 제시하였다.

---

16) REST(Representational State Transfer)ful 데이터는 다른 운영체제(OS), 웹브라우저 등의 소프트웨어를 사용하는 데이터 요청자를 대응하기 위한 데이터를 의미함.

17) DBMS(DataBase Management System)는 DB를 추가-수정-삭제할 수 있는 정보시스템을 의미함.

18) ERD(Entity-Relationship Diagram)는 DB의 구조를 모델링한 방법을 의미함.



〈그림 1-1〉 연구 수행 체계도

제2장

빅데이터의 개요 및 국내 현황

1. 빅데이터 개요

가. 빅데이터의 출현 배경

빅데이터는 사회 전반적으로 다양한 이해와 의미를 갖고 있는데, 단어적 의미로는 현재 다룰 수 있는 범위를 초과하거나 양적으로 다루기 어려운 데이터를 말하며 데이터의 수집-저장-분석-(시각화)-예측을 적시적으로 처리하기 어려운 형태나 데이터를 의미하기도 한다.

빅데이터가 등장한 가장 결정적인 이유는 저장장치의 혁명이었다. PC나 노트북 등 컴퓨터를 사용할 줄 아는 사람이면 누구나 문서, 이미지, 음악, 동영상 파일들이 PC나 노트북 내부에 설치되어 있는 하드디스크(HDD)에 저장된다는 사실을 알고 있다. 그런데 하드디스크(HDD)의 용량이 불과 10년 전인 2007년에는 1TB[19], 20년 전인 1997년에는 6GB 수준이었다.[20] 현재(2017년 8월 기준) 시장에는 12TB 용량의 하드디스크(HDD)와 8TB 용량의 플래시 메모리 형태의 저장매체가 유통되고 있다. 하드디스크(HDD)는 자기를 띤 원형의 디스크를 고속으로 회전시켜 데이터를 기록할 수 있는 공간을 찾아서 데이터를 기록하고 데이터가 존재하는 공간을 찾아서 읽어내는 다소 비효율적인 구조적인 특성이 있다. 이러한 구조적인 제약으로 빠른 IT기술의 발전에도 불구하고 데이터 저장용량과 입출력 속도는 한계에 다다랐다. 반면 1980~90년대 슈퍼컴퓨터, 특수목적용으로 사용되었던 플래시 메모

19) 2007년 하반기에 1TB 용량의 하드디스크(HDD)가 출시되었으나 당시에는 상대적으로 가격이 저렴한 500GB 용량의 하드디스크(HDD)가 시장을 주도함.

20) 1997년 3월 1일자 조선일보에 "6기가급 하드디스크 출시 붐"이라는 기사가 실렸다. 기사 내용에 "가격도 40만원선으로 1.6~2.5GB 하드디스크가 25만원선임을 감안할 때 충분한 경쟁력을"이 포함되었다.

리를 활용한 SSD(Solid-State Drive)가 2012년 이후 본격적으로 보급되기 시작하였다. 구조적인 데이터 병목이 해결되고 기술개발을 통한 용량 대비 단가 하락으로 데이터 입출력 속도가 크게 개선되어 빅데이터 시대를 열었다. 이미 컴퓨팅 처리속도와 네트워크 전송기술은 크게 앞서고 있었기 때문에 이러한 저장장치의 혁명은 현실과 가상세계의 차이를 빠르게 줄여나갔고 나아가 새로운 서비스 형태의 수익 모델을 만들어냈다. 구체적으로 사용하는 만큼 지불하는 클라우드 컴퓨팅 서비스가 등장했고 이를 바탕으로 정보서비스나 정보시스템의 진입장벽이 낮아져 카카오톡 등 카카오 서비스, 인스타그램, 핀터레스트, 페이스북 등 SNS와 모바일 서비스가 보편화되었다. 또한, 수많은 데이터를 생산하고 저장 및 활용할 수 있는 기회가 생겼고 물리적인 설치 장소 및 운영 공간 없이 이전 기술이나 방법론에 비해서 인적, 물적 자원의 초기 비용이 상대적으로 감소하여 많은 서비스와 데이터를 양산하게 되었다. 이는 「나. 빅데이터의 판단기준」에서 언급할 빅데이터의 3가지 특징을 만족시킬 수 있는 시발점이 되었다. 사회 전 분야에 걸쳐 데이터 서비스와 제공되는 데이터의 양적 증가와 더불어 데이터의 이동 및 처리 속도가 개선되어 연구자들은 다양한 가능성을 데이터 서비스와 데이터에서 찾을 수 있었고 이를 빅데이터라고 부르기 시작했다.

최근 이러한 빅데이터의 수집-저장-분석-(시각화)-예측을 적시적으로 처리하기 위한 방법으로 수집-저장 기술과 분석-시각화 기술, 분석-예측 기술, 수집-저장-분석-(시각화)-예측 기술이 발전하게 되었다. 특히 빅데이터를 빠르게 분석하고 검색할 수 있는 인공지능(AI)과 데이터 마이닝 분야가 발전하였고 우리나라에서는 2016년 구글(Google) 딥마인드(DeepMind Technologies Limited)[21]의 알파고(AlphaGo)와 이세돌 9단의 대국[22]에서 알파고가 4승 1패로 이세돌을 이김으로써 깊은 인상을 남겼다. 수학, 통계, 계산, 알고리즘과 자료구조, 프로그래밍, 병렬 및 분산 컴퓨팅, 소프트웨어 공학, 시스템 아키텍처, 통신 및 네트워크, 데이터베이스(DBMS), 머신러닝(Machine Learning), 딥러닝(Deep Learning), 자연어처리(Natural Language Processing), 로봇, 그래픽스, 휴먼 컴퓨팅, 가상현실, 생물정보, 인지과학, 계산물리학, 수치해석학, 암호학, 정보보안 등 다양한 컴퓨

21) 2014년 구글(Google)이 인수한 영국의 인공지능 기술회사(2010년 설립), 현재 알파벳 그룹 계열사.

22) AlphaGo versus Lee Sedul(Google DeepMind Challenge Match), 2016년 3월 9일부터 15일까지 서울에서 총 5회 대국함.

터 과학과 공학 기술을 중심으로 빅데이터를 지원하고 있다. 이를 활용해 새로운 부가가치를 창출하고자 경영공학, 금융공학, 생명공학, 의공학 등 융·복합 학문이나 산업분야가 발전하면서 데이터의 양이 더욱 급증하고 있다. 한국EMC와 IDC는 2020년 전 세계 디지털 데이터양이 44조GB로 예측되며 이중 1.9%에 해당하는 약 8,470억GB가 우리나라에서 생산될 것이라고 전망했다.[23] 따라서 빅데이터의 키워드가 사라졌을 뿐 여전히 데이터는 존재하고 앞으로도 더 증가하고 다양해질 것이기 때문에 빅데이터에 대해서 판단하고 준비해야 한다.

나. 빅데이터의 판단기준

2001년 메타그룹(현재 가트너)의 더그 레이니(Doug Laney)는 보고서[24]에서 빅데이터를 데이터의 양(Volume), 데이터의 속도(Velocity), 데이터의 다양성(Variety) 개념으로 표현하고 이를 관리할 수 있는 방안을 제시하였다. 이후 관련 연구에서 정확성(Veracity)이나 가치(Value) 등을 추가하여 재해석하기도 하였는데, 본 연구에서는 빅데이터의 개념에 대해 데이터의 양, 속도, 다양성의 3가지 측면에서 그 특징을 살펴보았다.

1) 데이터의 양(Volume)

연구에 활용되던 데이터가 기존에 활용되던 데이터로서 시간적, 공간적, 질적 혹은 다른 이유로 데이터가 양적으로 늘어난 경우이며, 예를 들면 시간적인 해상도를 높이기 위해서 측정 주기를 짧게 하거나 공간적인 해상도를 높이기 위해서 측정 장소 등 설치 위치의 밀도를 높이는 일, 또는 대기질 측정을 미세먼지(PM_{10})만 측정하다가 초미세먼지($PM_{2.5}$)를 추가로 측정하는 것이 있다. 한편 해당 분야에서 활용되지 않았던 데이터나 다른 분야의 데이터를 해당 분야 연구에서 활용함으로써 데이터가 양적으로 늘어나는 경우도 있다. 초미세먼지 연구에서 $PM_{2.5}$만 활용하다가 $PM_{1.0}$을 추가적으로 활용한다거나 미세먼지 데이터와 호흡기

23) https://korea.emc.com/about/news/press/2014/20140612.htm.

24) http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

질병 발생 데이터 간의 상관관계를 지역별로 활용하는 사례가 여기에 해당된다. 이와 같이 검토·분석해야 할 데이터의 양이 증가하여 수집-저장-분석-(시각화)-예측되는 데이터가 양이 증가하는 것을 의미한다.

2) 데이터의 이동속도(Velocity)

데이터 이동속도는 입출력 속도라고도 표현하는데 1)의 사례에서 보듯이 데이터의 양이 늘면 한 서버 단위에서 처리하기에는 저장이 어렵거나 저장이 되지 않고, 데이터베이스(DBMS)를 활용하기 어렵거나 활용이 불가능한 상황이 발생한다.[25] 또한 데이터의 개수가 늘어 일반적으로 기존의 PC나 서버에서 처리하기 어려워 연구자가 직접 처리하기 어려운 경우, 데이터의 흐름이나 처리속도가 느려질 수밖에 없다. 예를 들어 〈표 2-1〉에서 보면 일반적으로 연구자가 활용하는 PC가 MS 윈도우를 사용한다면 최대 파일 개수만큼 저장할 수 있다. 그런데, 실제로 한 폴더에 파일이 2,000개만 있어도 윈도우 탐색기의 속도가 현저히 느려지는 현상을 확인할 수 있다. 실내 대기질 연구를 하고자 연구자가 대기질 센서를 사무공간에 설치하여 4개의 센서에서 5초마다 1KB의 데이터를 저장한다고 하면 하루도 지나지 않아 연구자는 파일 목록을 확인하는 데도 곤란을 겪을 수밖에 없다.[26] 결국 연구자는 이 상황을 개선하기 위해서 데이터베이스(DBMS)를 도입하게 되는데, 한국환경공단의 대기질 측정소는 336개소이고 저장하는 데이터 종류도 많다. 더욱이 초당 1GB 이상의 데이터를 저장하는 스위스 제네바에 위치한 대형 강입자 가속기(Large Hadron Collider)라면 기존의 방법론으로는 수집-저장이 불가능하다. 대형 강입자 가속기가 생산하는 데이터를 단순 계산해도 하루 80TB 이상의 데이터를 저장하고 실제로 1,893TB의 데이터를 전송하는 것은[27] 기존 방법으로는 불가능하다. 이와 같이 결국 데이터의 양이 증가하여도 이동속도나 처리속도가 비슷하거나 개선됨을 의미한다.

---

25) 따라서 연구목적에 따라 기존의 파일시스템과는 다른 HDFS와 같은 빅데이터용 파일시스템을 사용하고 Hadoop과 같은 빅데이터 처리용 소프트웨어를 활용함. 또한, 기존의 DBMS와 다른 형태의 데이터를 저장하거나 고속의 대용량 데이터를 처리하기 위해서 NoSQL DBMS를 사용함.

26) 1분이면 48개, 1시간이면 2,880개, 하루면 6만 9,120개의 파일이 생성된다.

27) 대시보드(http://dashboard.cern.ch)로 확인한 2017년 6월 30일 0시부터 7월 1일 0시까지 전송량임.

〈표 2-1〉 윈도우 파일시스템 파일 개수 및 용량 제한

윈도우 파일시스템	한 볼륨에 넣을 수 있는 최대 파일 개수	파일 크기 제한
FAT32	4,194,304개	약 4GB
NTFS	4,294,967,295개	약 16TB

3) 데이터의 다양성(Variety)

연구과정에 활용되는 데이터의 유형이 변하여 1의 사례에서 보듯이 데이터가 단순히 양만 느는 것이 아니라 관심의 범위가 넓어져 기존에 활용하지 않았던 데이터들을 연구에 활용한다. 예를 들어, NoSQL과 같이 JSON 형태의 데이터나 XML 문서 형태의 데이터도 데이터베이스(DBMS)로 활용이 가능하지만 기존의 DB 활용방법으로는 불가능하다. 활용가능 데이터의 범위가 엑셀이나 데이터베이스(DBMS)에 저장된 정형화된 데이터들이었다면 정형뿐만 아니라 비정형 데이터도 포함하는 다양한 데이터들을 활용함을 의미한다. 데이터의 형태나 속성에 구애받지 않는다. 예를 들어, 환경 분야 특정 주제에 대해서 인터넷 웹 검색결과와 환경백서 문서 파일의 내용과의 연관성을 찾아보거나, 온라인 포털에서의 시의성을 분석하는 것은 이전의 방법론으로는 불가능한 방법이다. 필요한 모델이나 알고리즘이 포함된 라이브러리를 알면 굳이 내가 그 연구방법론을 모르더라도 손쉽게 재생할 수도 있고 깃허브(이하 'Github')[28]와 같은 온라인 공유공간을 이용해 나의 모델이나 알고리즘, 라이브러리를 공개하고 마치 같은 공간에서 연구하는 것과 같은 다양한 형태의 연구 활동도 가능하다.

앞서 살펴본 특징들은 시점의 차이일 뿐 사회 전반적으로 겪고 있는 현실이고 빅데이터라는 키워드 대신 기계학습(Machine Learning, 이하 'ML')[29]이나 심층학습(Deep Learning,

28) 깃허브(Github, http://www.github.com)란 온라인 코드 및 데이터 공유 공간이며 특정시점으로 코드 및 데이터를 복원할 수 있는 서비스임.

29) 기계학습, 또는 머신러닝(ML: Machine Learning)이란 인공지능의 한 분야로 컴퓨터가 스스로 또는 수동으로 학습할 수 있도록 컴퓨팅 알고리즘과 기술을 개발하는 분야를 말함, 더 넓은 의미로는 활용하는 분야도 포함함.

이하 'DL')[30], 혹은 데이터 트랜스포메이션(Data Transformation)[31], 데이터 등 좀 더 세분화되어 주변에 존재한다. 이에 기존 방법론으로는 인적, 물적, 시간적, 공간적으로 제약이 있는 빅데이터가 우리나라에서는 어떻게 자리 잡고 있는지, 더 나아가 환경 분야에는 어떻게 적용되고 있는지 살펴볼 필요가 있다. 특히 환경 분야는 다양한 연구주제와 데이터에 두루 걸쳐있고, 특히 산업이나 생활과 밀접한 연관성이 있어 실시간 또는 주기적인 데이터도 많다. 복합적인 학문이기 때문에 다양한 데이터와 그 연구방법론에 대한 이해가 필요하고 그 데이터의 양과 이동속도, 다양성은 빅데이터로서 관리되어야 하므로 체계적인 접근이 필요하다. 이에 국내 빅데이터 서비스를 살펴보고자 한다.

2. 국내 빅데이터 서비스 현황

가. 정부 주도형 빅데이터 서비스

국내 빅데이터 서비스는 『정보화촉진 기본계획』[32]을 기반으로 『스마트 국가 구현을 위한 빅데이터 마스터플랜』(2012)과 정부3.0 전략을 통해 정부 주도로 추진되어 왔으며 이는 세계 주요 국가도 우리와 비슷한 여건이다. 「공공데이터법」(공공데이터의 제공 및 이용 활성화에 관한 법률)[33]을 보면, 법령에 명시된 공공기관은 해당 공공기관이 보유 관리하는 공공데이터를 검토 및 조치 후 국민에게 제공해야 하며(제2조, 제17조) 해당 공공기관은 해당 공공기관의 소관 공공데이터 목록을 행정안전부장관에게 등록하고 행정안전부장관은 공공데이터 포털(https://www.data.go.kr, 이하 생략)에 공공데이터 목록에 관한 정보를 그 내용별, 형태별, 이용대상별 등 이용에 용이하게 분류하여 관리 제공하도록 규정되어 있다(제18조). 이에 공공데이터포털에서는 국토관리, 보건의료, 재난안전, 교통물류, 환경기상 등 16개 데이터 카테고리와 파일데이터, 오픈API, 표준데이터 등의 데이터셋을 제공한다.

---

30) 심층학습, 또는 딥러닝(DL: Deep Learning)이란 기계학습의 한 분야로 여러 비선형 변환기법의 조합을 통해 대량, 혹은 복잡한 데이터의 핵심적인 내용을 추출하거나 기능을 요약하는 분야를 말함.

31) 데이터 트랜스포메이션(Data Transformation)이란 모든 부가가치의 IT 산업화를 뜻하는 IT 트랜스포메이션(IT Transformation)처럼 부가가치가 데이터 산업화의 변화를 나타냄.

32) 「정보화촉진 기본법」에 따라 1996년부터 5년마다 확정, 2017년 현재 5차(2013~2017년)까지 수립.

33) 「공공데이터법」(공공데이터의 제공 및 이용 활성화에 관한 법률)은 법률 제14839호로 2017.7.26. 시행.

공공데이터포털 외에 대표적인 공공 빅데이터 서비스로는 통계청에서 제공하는 국가통계포털(KOSIS), 통계지리정보서비스(SGIS), 마이크로데이터 통합서비스(MDIS), 국가지표체계, 통계분류포털과 국토교통부에서 제공하는 국토교통 정보시스템, 행정안전부에서 제공하는 재난안전데이터포털, 환경부 및 그 산하기관에서 제공하는 에어코리아, 기상자료개방포털, 환경공간정보시스템, 국가상수도정보시스템, 환경통계포털, 서울특별시와 경기도에서 제공하는 서비스들이 있다(표 2-2 참조). 정부 부처와 지자체에서 운영하는 빅데이터 서비스 모델은 공공데이터포털과 같이 부문별로 제공하는 서비스를 모아서 수요 축으로 재분류하는 형태와 국가통계포털과 같이 통계 분류체계를 활용해 재분류하는 형태로 구분지을 수 있다. 민간에서는 SK텔레콤에서 제공하는 빅데이터 허브(Big Data Hub)[34]가 그 대표적인 사례로, 각 카드사와 통신사 등 상용 서비스 기업은 수익 모델을 반영하여 이용자 데이터 및 그 통계를 데이터를 판매하고 있으며 통계청의 마이크로데이터 통합서비스, 국민건강보험공단의 건강보험자료 공유서비스 등 공공기관에서 제공하는 공공서비스도 일부 유료로 제공하는 서비스도 있다.[35]

34) https://www.bigdatahub.co.kr/index.do.

35) 공공 부문의 빅데이터 서비스도 경우에 따라서는 운전자금을 마련할 수준의 수익 모델이 필요함.

〈표 2-2〉 주요 공공 빅데이터 서비스 현황

구 분	빅데이터 서비스명(URL)
행정안전부	· 공공데이터포털(https://www.data.go.kr)
	· 재난안전데이터포털(https://data.mpss.go.kr/Portal_new)
통계청	· 국가통계포털(http://kosis.kr)
	· 통계지리정보서비스(https://sgis.kostat.go.kr)
	· 마이크로데이터 통합서비스(https://mdis.kostat.go.kr)
	· e-나라지표(http://www.index.go.kr) 등 국가지표체계
	· 통계분류포털(https://kssc.kostat.go.kr)
환경부	· 에어코리아(http://www.airkorea.or.kr)
	· 기상기후 빅데이터 분석 플랫폼(https://bd.kma.go.kr)
	· 기상자료개방포털(https://data.kma.go.kr)
	· 환경공간정보서비스(https://egis.me.go.kr)
	· 물환경정보시스템(http://water.nier.go.kr)
	· 실시간수질정보시스템(http://www.koreawqi.go.kr)
	· 수질오염방제정보시스템(http://www.waterkorea.or.kr)
	· 국가상수도정보시스템(http://www.waternow.go.kr)
	· 국가하수도정보시스템(https://www.hasudoinfo.or.kr)
	· 토양지하수정보시스템(http://sgis.nier.go.kr)
	· 국가 대기오염물질 배출량 서비스(http://airemiss.nier.go.kr)
	· 실내공기질 자료공개 서비스(http://info.inair.or.kr)
	· 화학물질정보시스템(http://ncis.nier.go.kr) / 화학물질정보처리시스템
	· 국가소음정보시스템(http://www.noiseinfo.or.kr)
	· 환경통계포털(http://stat.me.go.kr)
국토교통부	· 국토교통 정보시스템(http://www.molit.go.kr/network)
.....	
서울특별시	· 서울 열린데이터 광장(http://data.seoul.go.kr)
	· 서울 통계(http://stat.seoul.go.kr)
	· 서울연구데이터서비스(http://data.si.re.kr)
경기도	· 경기데이터드림(http://data.gg.go.kr)
	· 경기통계(http://stat.gg.go.kr)
	· 지자체별 공공데이터포털(광장) : 수원, 성남, 오산, 부천, 화성, 안양, 안산

나. 공공데이터포털의 환경 분야 빅데이터 수요

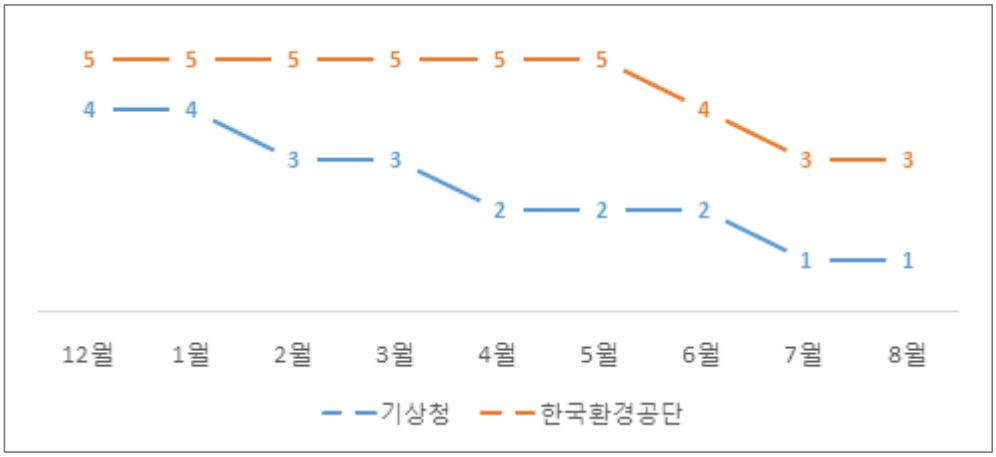
국내에서는 빅데이터가 공공데이터포털을 중심으로 각 부처별, 기관별로 운영되고 있어 공공데이터포털에서 제공하는 활용 신청 통계자료를 검토하였으며 연구자들의 환경 분야 빅데이터 수요는 한국환경정책·평가연구원의 연구자를 대상으로 설문조사하였다.

우선 2011년부터 2016년 12월까지 「공공데이터 활용 신청 TOP 20」[36] 중 「OPENAPI 활용 신청 TOP 20」을 보면 환경 분야 빅데이터로 판단되는 데이터는 기상청 (신)동네예보 정보조회서비스(누적 4위, 2,872건)와 한국환경공단의 대기오염정보 조회 서비스(누적 5위, 2,435건), 측정소정보 조회 서비스(누적 16위, 327건 이상)이고 나머지는 관광, 교통, 우편번호(주소), 부동산, 병원 관련 데이터이다. 특히 2016년 기준 기상청의 (신)동네예보정보조회서비스는 2,804건, 한국환경공단의 대기오염정보 조회 서비스는 1,704건, 측정소정보 조회 서비스는 327건으로 지난 한 해 동안 1, 2위 수준으로 전체적으로 해당 데이터에 대한 수요가 높았던 것을 확인할 수 있었다. 2017년 3월부터 공공데이터포털에서 순위 정보만 제공하는데, 제공된 데이터만 볼 경우에도 활용 신청이 지속적으로 늘어 이들 데이터에 대한 관심이 높다고 판단된다(그림 2-1 참조).

「매월 활용 신청 TOP 10」과 「2017년 누적 TOP 20」에서 특이한 점은 기상청의 경우 5월, 6월에 생활기상지수조회와 7월, 8월에 지상관측자료 월값(월보) 정보조회서비스 활용 신청이 많았고 한국환경공단의 경우 데이터의 특성상 지속적으로 측정소정보 조회 서비스의 활용 신청이 많았다는 점이다. 생활기상지수조회는 식중독지수, 체감온도, 열지수, 불쾌지수, 동파가능지수, 자외선지수, 대기오염확산지수를 제공하고 지상관측자료 월값(월보) 정보조회서비스는 지상관측장비 및 항공기상관측장비가 관측한 지상관측자료 및 항공기상관측 자료의 일값과 월값 정보를 제공하는 것으로 계절이 바뀜에 따라 생활과 관련이 있는 데이터에 대한 수요가 많은 것으로 해석할 수 있다.

---

36) 공공데이터포털 공지사항을 통해 2014년 4월부터 매월 공공데이터 활용 신청 TOP 10, 당해 연도 해당 월까지 누적 공공데이터 활용 신청 TOP 20와 2011년부터 해당 월까지 누적 공공데이터 활용 신청 TOP 20가 공개됨, 단, 2013년은 해당 년의 TOP 10만, 수치는 2017년 2월까지만, TOP 20는 2016년 6월부터 공개됨.



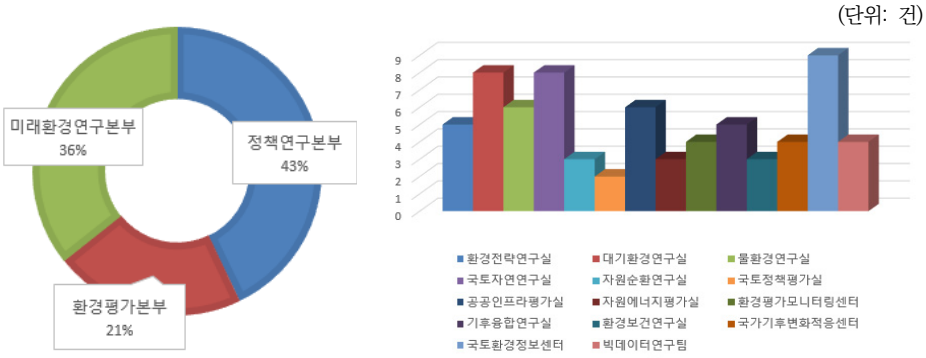
〈그림 2-1〉 공공데이터포털 데이터 활용 신청 순위(2016년 12월~2017년 9월)

다. 환경 분야 연구자의 빅데이터 수요

한국환경정책·평가연구원은 우리나라의 환경정책 수립을 위한 연구, 환경영향평가서 검토 및 이와 관련된 연구 활동을 수행하는 공공기관으로 환경 분야 빅데이터 수요 검증을 위해 적합하다고 판단하였다. 이에 연구 활동을 수행하는 연구자 220명을 대상으로 무작위 설문을 실시하여 70명이 응답하였다. 본 설문의 목적은 연구 활동에 디지털 데이터 활용 여부 및 어떤 디지털 데이터를 활용할지 탐색하고자 함이며, 특히 공공데이터포털 활용 신청 대비 기상기후 또는 대기질 데이터의 수요가 있는지를 확인하고자 함이다. 빠른 답변을 위해 총 7개 문항으로 작성하였고 모든 항목은 중복답변이 가능하도록 설정하였으며 설문은 네이버 오피스 폼[37]을 이용한 온라인 설문으로 작성하였다(부록 I 참조). 또한 설문결과는 평균 이상의 결과와 기타 의견에 대해서 언급하겠다.

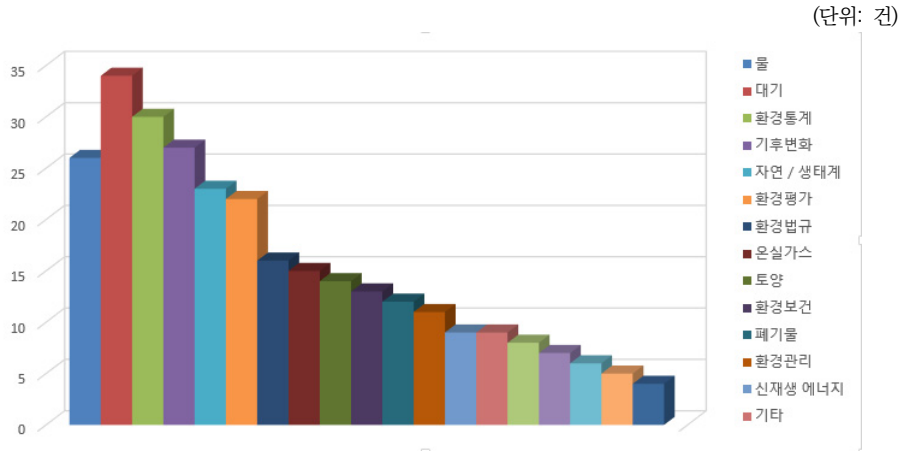
우선 설문 응답자는 〈그림 2-2〉와 같으며 대기환경연구실, 국토자연연구실, 국토환경정보센터 등의 답변이 높은 것으로 나타났으며 미응답 부서는 없었다.

37) 네이버 오피스 폼(https://office.naver.com).



〈그림 2-2〉 설문 응답자 분포

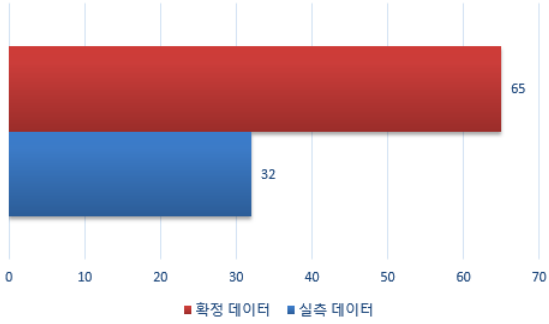
첫 번째, 환경 분야 중 어떤 데이터를 활용하는지에 대한 질문에 대해서는 대기 데이터 34명(48%), 환경통계 데이터 30명(42%), 기후변화 데이터 27명(38%), 물 데이터 26명(37%)이 사용한다고 응답하였다. 기타 의견으로는 환경뉴스나 환경보고서 등 환경 관련 비정형 데이터와 환경지출비용, 인구, 사망률, 발생률, 유병률 등 보건자료 외에도 특허, SNS 데이터를 활용한다고 응답하였다. 많이 사용되는 데이터 이외에 비정형 데이터와 타 분야 데이터에 대한 검토도 필요하다고 판단된다(그림 2-3 참조).



〈그림 2-3〉 연구 활용 데이터의 세부 분야

두 번째, 센서 등 실제 측정 데이터와 데이터 생산자가 확정한 데이터 중 어느 것을 많이 활용하는지에 대한 질문에는 실제 측정한 데이터보다 생산자가 확정한 데이터를 연구에 많이 활용하는 것으로 나타났다. 기타 의견으로는 연구자 모델링 수행 후 예측치를 활용한다고 답하여 연구과정에서 생산되는 데이터에 대한 고려가 필요하다고 판단된다(그림 2-4 참조).

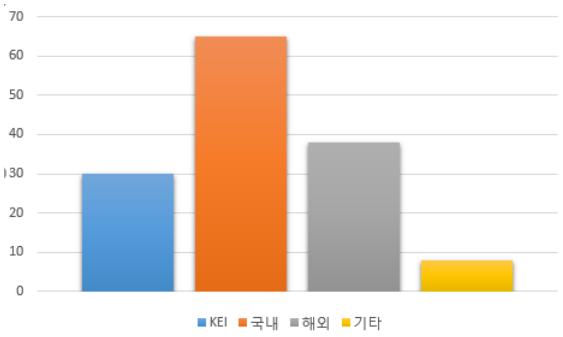
(단위: 건)



〈그림 2-4〉 연구 활용 데이터의 속성

세 번째로는 연구에 활용하는 데이터가 한국환경정책·평가연구원 내부에 있는지, 가령 외부에 있다면 국내에 있는지, 해외에 있는지, 그 외 알고 있는 데이터 및 그 위치에 대한 설문을 진행하였다. 65명이 한국환경정책·평가연구원이 아닌 우리나라에 있다고 답변하였고 38명이 해외에 있는 데이터를 활용하는 것으로 나타났다. 한국환경정책·평가연구원 내부 데이터의 활용에 대해서는 30명이 활용한다고 응답하여 상대적으로 외부 데이터 활용률이 높은 것으로 나타나 데이터 통합관리체계 구축 등 대응책 마련이 필요하다고 판단된다. 기타 의견으로 드론 등을 이용해 직접 제작한다는 답변이 있어 자체 생산 데이터에 대한 체계적인 관리와 적극적인 공유정책도 대응책 마련에 포함해야 할 것으로 판단된다(그림 2-5 참조).

(단위: 건)



〈그림 2-5〉 연구 수요 데이터의 위치

네 번째로, 외부의 어떤 빅데이터 서비스를 활용하는지에 대한 질문에는 공공데이터포털 37명, 국가통계포털 등 통계청 32명, 환경부 환경통계포털 31명, 기상자료개방포털, 기후정보포털 등 기상청이 29명, 한국환경공단 에어코리아가 26명 순으로 나타났다. 기타 의견으로 IEA[38], IIASA[39], World Bank Open Data[40], OECD[41], USGS[42] 등 해외 기관과 한국지질자연연구원, 국토지리정보원, 국립해양조사원, 환경부 환경공간정보서비스, 환경부, 농업 관련 데이터와 자료구축 기관을 소개 받아 직접 전달받는다는 의견이 있었다. 이에 빅데이터 서비스에 대한 메타데이터 구축이 필요하다고 판단된다(조사 항목은 부록 Ⅰ 참조).

다섯 번째는 데이터 활용 시 사용하는 소프트웨어에 대한 질문으로 엑셀이 65명, R이 29명, ArcGIS가 28명, SPSS가 19명, QGIS가 16명이었다. 언급할 만한 기타 의견은 없었고 R, QGIS 외 오픈소스로 파이선(Python)이 11명이 있었고 분석 및 시각화 소프트웨어인 테블로(Tableau)는 7명이 사용한다고 응답하였다. 활용 가능한 라이브러리가 많고 사용이 다소 쉬운 R과 파이선은 연구자 스스로 데이터를 수집-저장-분석-(시각화)-예측하는 데 사용할 수 있는 최적의 소프트웨어로 다음 설문 시 더 많이 사용하기를 기대한다(조사 항목

38) https://www.iea.org/statistics/, http://data.iea.org.
39) http://www.iiasa.ac.at/web/home/research/modelsData/models-tools-data.html.
40) World Bank Open Data(https://data.worldbank.org).
41) https://data.oecd.org.
42) https://data.usgs.gov.

은 부록 Ⅰ 참조).

여섯 번째 질문인 데이터를 연구에 활용하기 어려운 이유에 대해서는 필요한 데이터가 없거나 필요한 데이터가 어디 있는지 몰라서 활용하지 못한다는 답변이 각각 30명, 29명으로 나타났다. 절차적인 애로사항으로 신청절차가 복잡하거나 데이터를 수집하기 어렵다는 답변이 24명, 25명이었다. 사용방법에 대한 어려움으로 데이터 전처리가 곤란하거나 데이터 관련 소프트웨어를 사용할 줄 모른다가 각각 21명, 21명이었다. 기타 답변으로는 어려움이 없거나 또는 어려움은 없으나 전처리 시 시간이 많이 걸린다는 답변이 있었다. 연구에 필요한 데이터가 어디에 있는지 모르는 경우는 데이터 통합관리체계 마련과 적극적인 홍보로 해결되리라 기대하고 절차적인 애로사항과 사용방법에 대한 어려움은 연구 활용사례 발굴 및 사례기반 학습으로 해결할 수 있을 것이다.

마지막의 데이터 기반 연구를 수행하기 위해서 무엇이 필요하냐는 질문에는 다음과 같이 다양한 답변이 작성되었다.

1) 소프트웨어 구입비용이 비싸거나 오픈소스로 활용 가능한 소프트웨어가 있으나 이해와 사용상의 어려움
2) 활용할 수 있는 라이브러리나 모델을 모르고 자체 교육이 필요함
3) 데이터 활용을 위한 메타데이터가 필요하고 데이터 지원인력이나 조직이 필요함
4) 데이터 공유를 위한 유용한 서비스를 활용할 수 있었으면 함(정보보안 측면)
5) 따라서 위 문제들을 해결하기 위해서 조직 차원에서 대응방안을 마련해 주었으면 함

본 설문조사는 환경 분야 연구에 기상기후나 대기질 데이터의 우선순위를 파악하고자 함으로 위의 문제에 대해서 깊이 고민할 수는 없겠지만, 조직 차원에서 소프트웨어 구입과 활용, 교육에 대한 주기적인 수요조사를 바탕으로 데이터 기반구축 및 활용(안) 등 데이터 통합관리체계를 마련하여 체계적으로 데이터 수요에 대한 대응을 하고 단계별 접근전략을 통해 위와 같은 의견들이 반영될 수 있을 것으로 판단된다.

이상의 공공데이터포털 활용 신청 건수 검토와 연구자 설문조사를 통해 기상기후와 대기질 데이터가 중요하다고 판단되었고, 이중 본 연구범위를 고려하여 한국환경공단의 대기질 데이터를 중심으로 환경 분야 빅데이터 수집방법을 검토하였다.

3. 대기질 빅데이터 서비스

가. 공공데이터포털(행정안전부)

한국환경공단이 공공데이터포털을 통해 제공하는 대기질 빅데이터는 파일데이터 19건, 오픈API 7건이다. 먼저 파일데이터를 살펴보면, 도로 재비산먼지 측정 정보, 재활용가능자원 가격조사, 수도 관련 수질자료, 폐기물 재활용 관련 통계, 상수도 통계, 폐기물 배출 사업장 정보, 석면피해인정 통계, 영농폐기물조사, 음식물쓰레기 배출정보, 폐기물통계정보, 소음진동측정망 운영정보, 공공하수처리시설 운영 현황, 폐기물배출 및 처리현황, 폐기물감량정보, 상수도정보 통합DB, 하수도 보급현황, 운행자동차 전문정비자료, 운행자동차 저감장치 사업 현황, 환경부전기차충전소 급속충전기 위치정보 등이며 에어코리아를 통해 서비스하고 있는 대기질 데이터는 없다. 그렇지만 도로 재비산먼지 측정 정보를 조회해 보면 파일데이터 특성에 각각의 파일마다 업데이트 주기, 차기등록예정일, 비용부과 유무, 비용부과 기준 및 단위, 다운로드 횟수, 등록일, 수정일, 이용허락범위, 제공형태, URL, 설명, 기타 유의사항, 관리부서명, 관리부서 전화번호가 메타데이터로 등록되어 있다. 파일을 다운로드하여 압축파일을 열어보면, CSV 파일과 엑셀 파일이 들어 있고 각각의 파일에는 측정일자, 측정시간, 지역, 지역명, 도로명, 시작점, 종점, 기온, 습도, 재비산먼지 평균농도, 오염범례로 측정내용이 정리되어 있다. 이 부분은 데이터를 다운로드하기 전에 해당 데이터에 대한 메타데이터가 제공되거나 내용을 볼 수 있도록 개선이 필요하다.

오픈API는 대기오염정보 조회 서비스, 전기차충전소 운영현황, 오존황사 발생정보조회 서비스, 대기오염통계 서비스, 측정소정보 조회 서비스, 음식물쓰레기 배출정보, 공공기관 연계자료 서비스가 제공된다. 연구에 활용할 수 있는 대기질 데이터는 대기오염정보 조회 서비스, 대기오염통계 서비스, 측정소정보 조회 서비스이다. 모두 공공데이터포털에 가입 후 로그인하여 신청하면 된다. 공공데이터포털에 로그인하여 대기오염정보 조회 서비스를 선택하여 들어가면 활용 신청 버튼이 보이고 개발계정 신청 페이지로 연결된다. 시스템 유형은 '일반'(기본 선택되어 있음), 활용 목적은 '연구(논문 등)'를 선택하고 상세기능정보의 상세기능을 확인 후 선택한 후에 라이선스 표시 중 이용허락범위(저작자표시)를 확인하고 '동의합니다' 앞에 체크한 후에 신청을 클릭하면 바로 신청이 된다. 단, 신청은 인터넷 익스

플로러를 통해서만 가능하고[43] 신청 후 서비스 정보 중 일반 인증키를 확인하여도 해당 서비스에 연동되는 데 1시간 정도 걸린다고 하니 바로 사용은 어렵다. 해당 서비스에 연결되지 않는 경우 웹이나 전화로 처리요청을 할 수 있다. 'End Point' 등 서비스 정보를 자세히 보면 실제 서비스는 해당 기관에서 제공하고 있음을 확인할 수 있고 참고문서를 통해 데이터를 활용하기 위한 메타정보와 요청/ 응답 메시지 등 사용방법을 제공하고 있다. 상세기능정보를 통해 미리보기를 수행할 수 있으며, 키가 등록되지 않은 경우 'SERVICE KEY IS NOT REGISTERED ERROR.'와 같은 메시지를 확인할 수 있다. 서비스 정보 중 참고문서를 보면 'Airkorea 국가대기오염정보 OpenAPI 활용 가이드'라는 문서를 확인할 수 있고 문서에 본 빅데이터 서비스에 대한 메타데이터 및 연동방법이 담겨 있다. 이를 보고 데이터를 조회하면 되고 한국환경공단에서 제공하는 OpenAPI는 JSON과 XML 형태로 데이터를 받을 수 있다. 이에 대한 코드 작성이 필요하며 웹브라우저에서 확인하고자 하는 경우, 요청 메시지 명세에 항목구분 중 필수(1)로 표시된 항목을 빼고 데이터를 요청하면 오류화면이 나온다. 예를 들어, 문서에 요청/응답 메시지 예제를 보면 〈표 2-3〉과 같다.

〈표 2-3〉의 내용을 웹브라우저의 주소창에 그대로 입력하면 오류가 발생한다.

〈표 2-3〉 오픈API 데이터 요청 메시지(예)

http://openapi.airkorea.or.kr/openapi/services/rest/ArpltnInforInqireSvc/getMsrstnAcctoRltmMesureDnsty?stationName=종로구&dataTerm=month&pageNo=1&numOfRows=10&ServiceKey=서비스키&ver=1.3

서비스키를 서비스 정보의 일반 인증키(UTF-8)로 표시된 주소로 바꿔서 보내야 한다. 또한 '종로구'를 '아름동'으로 변경하고 주소 제일 뒤에 '(&_returnType=json'를 추가하면 최근 10개의 측정데이터를 받을 수 있다. 따라서 프로그래밍을 할 수 없는 경우 한 줄 한 줄에 대한 검증은 가능하나 OpenAPI를 사용하는 방법은 사실상 불가능하다. 〈표 2-4〉와 같은 응답 메시지를 연구자가 활용하기는 어렵다.

43) 2017년 7월 20일까지 확인한 사항임.

〈표 2-4〉 오픈API 데이터 응답 메시지(예)

{"list":[{"_returnType":"json","coGrade":"1","coValue":"0.5","dataTerm":"","dataTime":"2017-06-30 03:00","khaiGrade":"2","khaiValue":"52","mangName":"도시대기","no2Grade":"1","no2Value":"0.017","numOfRows":"10","o3Grade":"1","o3Value":"0.003","pageNo":"1","pm10Grade":"2","pm10Grade1h":"1","pm10Value":"29","pm10Value24":"32","pm25Grade":"2","pm25Grade1h":"2","pm25Value":"19","pm25Value24":"16","resultCode":"","resultMsg":"","rnum":0,"serviceKey":"","sidoName":"","so2Grade":"1","so2Value":"0.001","stationCode":"","stationName":"","totalCount":"","ver":""}(삭제)],"parm":{"_returnType":"json","coGrade":"","coValue":"","dataTerm":"MONTH","dataTime":"","khaiGrade":"","khaiValue":"","mangName":"","no2Grade":"","no2Value":"","numOfRows":"10","o3Grade":"","o3Value":"","pageNo":"1","pm10Grade":"","pm10Grade1h":"","pm10Value":"","pm10Value24":"","pm25Grade":"","pm25Grade1h":"","pm25Value":"","pm25Value24":"","resultCode":"","resultMsg":"","rnum":0,"serviceKey":"(삭제)","sidoName":"","so2Grade":"","so2Value":"","stationCode":"","stationName":"아름동","totalCount":"","ver":"1.3("},"ArpltnInforInqireSvcVo":{"_returnType":"json","coGrade":"","coValue":"","dataTerm":"MONTH","dataTime":"","khaiGrade":"","khaiValue":"","mangName":"","no2Grade":"","no2Value":"","numOfRows":"10","o3Grade":"","o3Value":"","pageNo":"1","pm10Grade":"","pm10Grade1h":"","pm10Value":"","pm10Value24":"","pm25Grade":"","pm25Grade1h":"","pm25Value":"","pm25Value24":"","resultCode":"","resultMsg":"","rnum":0,"serviceKey":"(삭제)","sidoName":"","so2Grade":"","so2Value":"","stationCode":"","stationName":"아름동","totalCount":"","ver":"1.3("},"totalCount":717}

공공데이터포털은 내가 필요로 하는 데이터가 어느 기관에서 제공하는지, 또한 그 기관에서 어떤 빅데이터 서비스를 제공하는지 확인하는 데 유용한 것은 물론 프로그래밍이 가능하다면 OpenAPI 연동 후 데이터 사용이 가능하다. 그럼 에어코리아에서는 어떤 서비스를 제공하는지 살펴보겠다.

나. 에어코리아(한국환경공단)

한국환경공단은 도시대기 측정망 272개소, 교외대기 측정망 22개소, 국가배경농도 측정망 3개소, 도로변대기 측정망 37개소 등[44]을 통해 우리나라에서 가장 많은 대기질 데이터를 보유하고 에어코리아를 통해 서비스하고 있다. 〈표 2-5〉에서와 같이 공공데이터포털과 에어코리아를 통해 동일한 데이터를 제공한다.

---

44) 공공데이터포털 OpenAPI를 통해 제공하는 측정소정보 조회 서비스 통계 기준임.

〈표 2-5〉 한국환경공단 대기질 서비스

데이터 구분	데이터 속성	에어코리아	공공데이터포털
대기오염정보조회 서비스	실시간	웹 조회	오픈API(실시간)
오존황사 발생정보조회 서비스	실시간	웹 조회	
대기오염통계 서비스	확정	웹 조회	
측정소정보 조회 서비스	확정	웹 조회	
공공기관연계자료 서비스		제공 안 함	

에어코리아는 측정망과 측정소 정보, 실시간 대기질 자료, 대기질 예·경보, 최종확정자료, 배움터와 고객의 소리 등 정보채널로 구성되어 있으며, 측정망과 측정소 정보는 참고하되 필요한 경우는 공공데이터포털 OpenAPI를 통해 연동하는 편이 유리하다.

실시간 대기질 자료로는 우리동네 대기 정보와 미세먼지 정보가 데이터를 모으는 데 유용하며, 실제로 미세먼지 정보가 대량의 데이터를 수집하는 데 효과적이다. 이 경우 웹서비스에서는 2017년 현재까지의 데이터만 조회할 수 있으나 실제로는 2008년부터 데이터를 조회할 수가 있어 필요한 경우 넓은 범위의 데이터를 연구에 활용할 수 있다. 확정데이터도 2014년부터 제공되므로 약 10년치 데이터를 연구에 활용할 수 있다고 판단할 수 있다. 데이터를 수집하는 방법은 다음과 같다.

1) 웹서비스와 웹페이지를 분석하고 시험용 URL를 만들어낸다.
2) 시험용 URL을 통한 결과가 확인되면 데이터 저장을 위한 데이터베이스(DBMS)를 구성한다.
3) 수집을 위한 자동화 코드를 작성한다.
4) 데이터 연결을 위한 R과 파이선용 코드, 데이터 활용을 위한 메타데이터를 작성한다.
5) 데이터를 DBMS에 담고 코드와 메타데이터를 확인한 후 공유한다.

〈그림 2-6〉과 같은 코드를 통해 웹페이지에서 조건을 클릭하는 것과 동일하게 데이터를 조회할 수 있다.

```
#url = r"http://www.airkorea.or.kr/pmRelayDay?itemCode=10008
strDateDiv = "strDateDiv=2"
searchDate = "searchDate=2006-01-01"
district = "district=02"
itemCode = "itemCode=10008"
searchDate_f = "searchDate_f=200601"

#일평균 PM10 실시간 자료조회
url = r"http://www.airkorea.or.kr/pmRelaySub?" \
        + strDateDiv + r"&" + searchDate + r"&" + district + r"&" + searchDate_f \
        + r"&itemCode=10007"
#url
df = pd.read_html(url, encoding="utf-8")
#df
df[0]
```

	지점ID	지점	1	2	3	4	5	6	7	8	...	22	23	24	25	26	27	28	29	30	31
0	도시대기	[서울]강남구	60	68	31	28	36	34	34	39	...	32	38	40	73	94	98	109	75	69	38
1	도시대기	[서울]강동구	64	74	31	33	43	46	44	47	...	39	45	62	81	113	115	111	81	76	40
2	도시대기	[서울]강북구	51	59	26	27	29	33	30	39	...	29	36	41	66	89	95	101	75	59	38
3	도시대기	[서울]강서구	78	74	36	31	42	49	43	46	...	40	44	53	91	114	116	130	92	75	55

〈그림 2-6〉 데이터 수집 코드와 조회결과

대기질 예·경보 중 대기질 예보는 이미지 데이터로 파일 다운로드 형태로 자동화하면 되는데, 데이터의 활용도가 다소 떨어진다. 대기질 경보 데이터는 기간을 넓혀서 한 번에 다운로드받거나 스케줄러와 자동화 코드를 병행하여 공공데이터포털 OpenAPI를 통해 연동할 수 있다.

통계정보 중에서는 측정소별 확정자료와 측정망·항목별 확정자료가 자동화하기 용이하고 대기환경 연월보와 최종 확정자료 다운로드는 목록화 후 파일 다운로드로 수집할 수 있으며, 국외대기오염현황은 연 1회 업데이트되기 때문에 변경 발생 시를 대비하여 알람 스케줄러를 설정하는 것이 좋다. 측정망·항목별 확정자료는 2014년 1월 1일부터 데이터 조회가 가능하다. 따라서 코드로 조회하는 것이 데이터를 확보하는 데 유리하다.

따라서 빅데이터 수집-저장 방법을 절차화하고 자동화 코드와 메타데이터, 데이터를 통합 관리할 수 있는 프레임워크로 제안하고자 한다.

제3장

환경 분야 빅데이터 수집방법

1. 환경 분야 빅데이터 수집절차

가. 수집절차의 도식화

앞서 살펴본 환경 분야 빅데이터 중 대기질 데이터의 수집방법에 대한 절차를 정리해보면 다음과 같다.

> 첫째, 수집하고자 하는 대상 데이터를 검토 및 선정하고 대상 데이터 제공자를 찾는다.
> 둘째, 데이터 제공자가 제공하는 데이터 형식을 파악하고 적절한 대응방법을 찾는다.
> 셋째, 데이터를 수집하고 활용 가능한 형태로 저장한다.

이를 풀어보면, 수집하고자 하는 데이터가 대기질 빅데이터라고 하자. 제일 먼저 공공데이터 포털에서 대기질을 검색한다. 파일데이터가 5건, 오픈API가 9건이 검색될 것이다. 그 다음은 메타데이터를 보면서 데이터 제공자를 찾는다. 파일데이터에서는 만족할 만한 데이터 제공자가 없을 것이고 오픈API에서는 지방자치단체와 한국환경공단이 확인된다. 따라서 전국 대기질 데이터를 모으려면 한국환경공단 오픈API를 신청하게 될 것이고 한국환경공단, 대기질 등의 키워드로 구글과 같은 검색엔진에서 별도의 빅데이터가 서비스되고 있는지 찾게 된다. 물론 해당 기관의 홈페이지를 접속하여 바로가기 등을 확인해도 된다. 그 다음은 해당 빅데이터 서비스나 웹서비스 등을 분석하여 빅데이터 수집-저장 자동화 코드를 작성하고 데이터베이스(DBMS)를 구성한다. 이 코드를 구동하면서 DBMS에 저장하기만 하면 데이터 수집절차가 진행된다. 이것을 프레임워크로 만들기 위해서는 절차를 다음과 같이 나누어 볼 수 있다.

1) 빅데이터 수집 요청서 또는 수요조사
2) 빅데이터 수집 검토서
3) 빅데이터 수집-저장 요구사항 정의서(명세서)
4) 빅데이터 수집-저장 상세 요구사항 정의서(명세서)
5) 데이터베이스(DBMS) 테이블 정의서(명세서)
6) ERD
7) 빅데이터 메타데이터 설명서
8) 빅데이터 수집-저장 프로그램 정의서(명세서)

위의 내용은 빅데이터 수집-저장을 위한 절차로 세부내용은 다음 단락에서 다루기로 한다.

나. 수집절차의 상세화

빅데이터 수집 요청서는 해당 데이터가 필요한 연구자가 작성한다. 필요한 데이터 정보(데이터명, 온라인-오프라인 여부, 파일 다운로드, 압축 여부, 메타데이터 존재 유무, 질의 가능 여부와 질의 정보 등)와 데이터 제공자 정보(데이터 제공자명, 담당자 연락처, 신청절차 유무, MOU 체결 유무 등)를 작성한다.

빅데이터 수집 검토서는 수집 요청서에 기재된 빅데이터 서비스를 검토하여 작성한다. 구체적으로 수요자 정보, 빅데이터 서비스 유형(웹 조회, 웹 다운로드, API 유무 등), 수집-저장할 데이터 유형, 메타데이터 확인 유무, 메타데이터 특이사항, 시험용 코드 작성 및 확인 유무 등이다.[45]

해당 데이터가 식별되면 빅데이터 수집-저장 요구사항 정의서(명세서)를 작성한다. 또한, 수집-저장 절차가 범용적인 활용이 가능하도록 수요 연구자와 협의하여 작성한다, 아래와 같이 간단하게 작성해도 된다(표 3-1 참조).

45) 데이터의 필요 유무는 데이터 변경통제위원회에서 심의함.

〈표 3-1〉 빅데이터 수집-저장 요구사항 정의서(예)

요구사항 번호	요구사항 내용	비 고
BD_REQ17_001	실시간 미세먼지 데이터가 관리되어야 한다.	에어코리아
BD_REQ17_002	에어코리아 실시간 자료조회 〉 미세먼지 정보 〉 PM_{10}	
BD_REQ17_003	에어코리아 실시간 자료조회 〉 미세먼지 정보 〉 $PM_{2.5}$	
BD_REQ17_004	에어코리아 실시간 자료조회 〉 미세먼지 정보 〉 금속성분	

그 다음으로는 빅데이터 수집-저장 상세 요구사항을 작성한다. 이 부분은 최대한 자세하게 확인하고 작성하여야 한다. 또한 요구사항 명세에 대한 추적이 가능해야 하고 수집-저장 자동화 코드 작성 후 체크리스트로 활용할 수 있어야 한다(표 3-2 참조).

〈표 3-2〉 빅데이터 수집-저장 상세 요구사항 정의서(예)

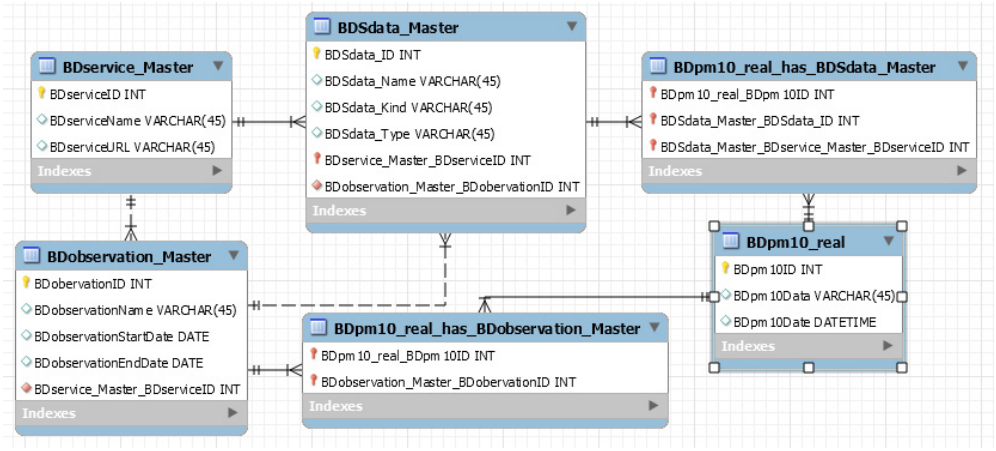
번호	요구사항 내용	요구사항 번호
00101	데이터베이스(DBMS)를 구성한다.	BD_REQ17_001
00102	데이터베이스(DBMS) 구성을 위한 테이블 명세서를 작성한다.	BD_REQ17_001
00103	데이터베이스(DBMS) 구성을 위한 ERD를 작성한다.	BD_REQ17_001
00201	데이터베이스(DBMS)에 측정망, 측정소, 시간 단위로 저장해야 한다.	BD_REQ17_002
00202	일평균 수치는 뷰테이블로 구성하며 일평균 조회결과를 검증한다.	BD_REQ17_002
00203	2012년부터 조회가 가능하면 2012년부터 수집-저장한다.	BD_REQ17_002
00301	데이터베이스(DBMS)에 측정망, 측정소, 시간 단위로 저장해야 한다.	BD_REQ17_003
00302	일평균 수치는 뷰테이블로 구성하며 일평균 조회결과를 검증한다.	BD_REQ17_003
00303	2012년부터 조회가 가능하면 2012년부터 수집-저장한다.	BD_REQ17_003
00401	데이터베이스(DBMS)에 측정소, 시간 단위로 저장해야 한다.	BD_REQ17_004
00402	과거부터 조회가 가능하면 해당 시점부터 수집-저장한다.	BD_REQ17_004

DBMS에 데이터를 저장하기 위해서 테이블 정의서를 작성하여 DBMS 스키마를 적용한다. 이력 관리를 하고자 할 때는 추가일시, 추가자, 수정일시, 수정자를 반영하거나 트리거로 구성하고 이 내용을 상세 요구사항에 추가한다(표 3-3 참조).

〈표 3-3〉 데이터베이스 테이블 정의서(예)

테이블명	요구사항 내용	비고
BDservice_Master	서비스ID, 서비스명, 서비스URL	00101
BDSdata_Master	데이터ID, 데이터명, 데이터단위, 데이터 속성, 서비스ID, 관측소ID	00101
BDobservation_Master	측정소ID, 측성소명, 데이터ID, 측정시작일자, 측정종료일자, 서비스ID	00101
BDpm10_real	PM10데이터ID, 측정값, 측정일시, 측정소ID, 데이터ID	00201
vwBDpm10_real	(SQL 쿼리로 작성)	00202

테이블 정의서를 작성하면 DBMS에 구조(Schema)를 구성하기 위한 ERD를 작성하고 검토한다. ERD 작성 시 MySQL Workbench와 같은 오픈소스 소프트웨어를 사용하면 편리하다(그림 3-1 참조).



〈그림 3-1〉 ERD(예)

ERD까지 작성되면 DBMS를 구축하고 자동화 코드를 운영한다. 주기적인 데이터 수집-저장이 필요한 경우는 스케줄러를 통해 정해진 시간에 데이터를 동기화하도록 한다. 스케줄러를 통해 동기화하는 경우, 데이터 중복에 대한 처리를 코드에 반영하는 것이 좋고 동기화 로그를 남기도록 한다.

수집-저장 자동화 코드가 정상 동작하면 빅데이터 메타데이터 설명서를 작성하고 연구자들에게 배포할 준비를 한다. 설명서에는 메타데이터 정보와 메타데이터 접근방법, 예시를 작성한다. 메타데이터 정보는 〈표 3-4〉에서와 같이 메타데이터명과 데이터의 출처, 속성, 유형, 주기, 단위, 위치 등 메타데이터 설명을 작성하고 메타데이터 접근방법은 〈표 3-5〉와 같이 서비스에 대한 주요 내용을 제시한다.

〈표 3-4〉 메타데이터 정보(예)

구 분	내 용	비고
메타데이터명	한국환경공단 미세먼지 실시간 데이터	PM_{10}
데이터 출처	한국환경공단 에어코리아 실시간자료조회 〉 미세먼지 정보 〉 PM_{10}	
데이터 속성	실시간 비확정	
데이터 유형	웹 조회결과 추출	
데이터 주기	시간	
데이터 단위	$\mu g/m^3$	
측정소 위치	TM 좌표	

〈표 3-5〉 메타데이터 접근방법(예)

구 분	내 용	신청 필요	비고
DBMS	☐ MariaDB ☑ PostgreSQL	○	테이블 명세서 참조
	- DB 명 : realtimepm10 - 서비스 호스트 : pm10.realtime.data.org - 서비스 IP : 11.22.33.44 - 서비스 포트 : 5432 - 접속 ID 및 비밀번호 : 신청 시 지정		
수집기	☐ R ☑ Python		
	- 코드 : www.github.com/metagetpm10 ※ 프로그램 명세서 참조		
오픈API	☐ SOAP ☑ REST	○	
	- 서비스 호스트 : pm10.realtime.data.org - 서비스 URI : /getdata - POST 메시지 1) datarange : all(전체), year(년), month(월) 2) dataresol : hour(시간), day(일평균) 3) fromdate : YYYYMMDDHH(시작 연월일시) 4) todate : YYYYMMDDHH(종료 연월일시) 5) location : all(전 지역) 6) location_x : TM X좌표 7) location_y : TM Y좌표		
응답 메시지	☐ JSON ☑ XML	○	
	※ 응답 메시지 설명서 참조		

마지막으로 〈표 3-6〉과 같이 작성한 자동화 코드에 대한 빅데이터 수집-저장 프로그램 정의서(명세서)가 작성되어야 이를 활용하는 연구자가 코드를 바로 활용하거나 수정하여 활용할 수 있다.

「가. 수집절차의 도식화」에서 언급된 1)부터 8)까지의 절차가 완료되어도 수정사항이 발생할 경우, 이력관리가 이루어져야 한다. 기술 등 여건변화가 모든 연구자에게 동일하게 적용되는 것이 아니기 때문에 다음 연구자에게는 선행 연구자의 수정사항 등에 대한 이해가 필요하다. 따라서 수집방법에 대한 절차의 운영관리와 수정사항 등에 대한 이력관리를 수행할 수 있는 빅데이터 관리체계가 필요하다.

〈표 3-6〉 프로그램 명세서(예)

구 분	내 용	비고
프로그램명	get_pm10_airkorea.py	
프로그램 설명	사용방법에 명시된 조건들을 활용하여 한국환경공단의 에어코리아에서 제공되는 미세먼지 실시간 데이터를 수집(get_pm10_data)하여 저장(put_pm10_to_csv)하는 프로그램임	
작성한 언어	☐ R ☑ Python	
사용방법	- 코드 제일 앞의 datarange, dataresol, fromdate, todate, location, location_x, location_y, filepath, filename을 설정한 후에 python get_pm10_airkorea.py와 같이 실행한다. - 변수 설명 1) datarange : all(전체), year(년), month(월) 2) dataresol : hour(시간), day(일평균) 3) fromdate : YYYYMMDDHH(시작 연월일시) 4) todate : YYYYMMDDHH(종료 연월일시) 5) location : all(전 지역) 6) location_x : TM X좌표 7) location_y : TM Y좌표 8) filepath : 파일 저장 위치 9) filename : 저장할 파일명	
활용 라이브러리	import urllib.request from urllib.request import urlretrieve import time from bs4 import BeautifulSoup import pandas as pd	

함수 설명	// 미세먼지 데이터 수집범위
	- get_pm10_data(datarange, dataresol, fromdate, todate, location, location_x, location_y)
	
	// 미세먼지 데이터 저장위치
	- put_pm10_to_csv(filepath, filename)	

2. 빅데이터 수집-저장 프레임워크

가. 빅데이터 관리체계의 필요성

한국환경공단이 제공하는 대기질 빅데이터를 공공데이터포털의 오픈API를 통해 조회하면 코드 몇 줄로 한 번에 모든 측정소 정보를 얻을 수 있다. 측정소명(stationname), 측정소 주소(addr), 설치년도(year), 관리기관명(oper), 측정소 이미지(photo), 측정소 전경(vrml), 측정소 지도이미지(map), 측정망(mangname), 측정항목(item), 위도(dmX), 경도(dmY) 등을 XML이나 JSON 형태로 조회할 수 있다(그림 3-2 참조).

```
</item>
<item>
<stationname>부평역</stationname>
<addr>인천 부평구 광장로 지하15(부평동)부평역 7번 출구</addr>
<year>2010</year>
<oper>인천광역시보건환경연구원</oper>
<photo>http://www.airkorea.or.kr/airkorea/station_photo/NAMIS/station
<vrml></vrml>
<map>http://www.airkorea.or.kr/airkorea/station_map/823634.gif</map>
<mangname>도로변대기</mangname>
<item>SO2, CO, O3, NO2, PM10, PM2.5</item>
<dmx>37.482874</dmx>
<dmy>126.704943</dmy>
</item>
<item>
```

〈그림 3-2〉 조회결과(예)

그런데 측정소 이미지와, 측정소 전경, 측정소 지도이미지는 에어코리아에서 제공하는 데이터에 비해 데이터도 부실하고 데이터에 부재에 따른 오류가 처리되어 있지 않다. 그러나 에어코리아 웹페이지 분석을 통해 각 측정소의 코드를 추정할 수 있고 이미지 링크 등을 보정할 수 있었다. 특히, 측정소의 코드 값과 공간정보(위도와 경도)는 데이터의 무결성(검증)과 공간정보 결합 등 활용성 측면에서 중요한 데이터이므로 한국환경공단에서 제공하는 대기질 메타데이터 DB 구축 시 꼭 필요한 항목이다.

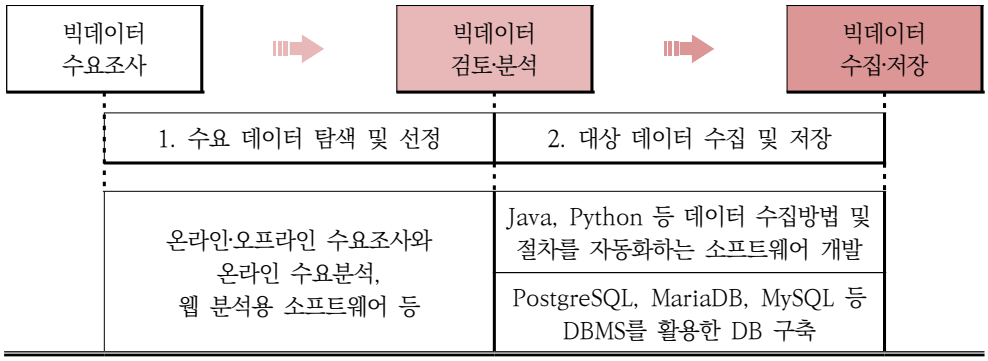
이처럼 연구 등 사용목적에 따라 다양하게 활용할 수 있도록 빅데이터 서비스의 서비스 데이터를 검토하고 서비스 유형을 분석하여 데이터 전처리를 포함한 데이터의 수집-저장

시 연구자 등 이용자에게 적합한 데이터 수집-저장 프레임워크와 더불어 일관된 데이터 관리체계가 필요하다.

나. 빅데이터 수집-저장 프레임워크

빅데이터 수집-저장 프레임워크란 한국환경공단의 에어코리아처럼 빅데이터 서비스에서 제공하는 데이터의 수집방법을 절차화한 것으로서, 그것을 수집한 데이터의 공유 및 활용을 위한 방법론을 말한다. 한국환경정책·평가연구원(KEI) 등 연구자들이 활용하는 데이터는 원본성과 무결성이 유지되어야 한다. 다시 말해, 연구 탐색-설계-실행-결과(보고서) 등 연구 전 과정에서 활용되는 데이터의 품질이 유지되어야 한다는 말이다.

빅데이터의 품질이 유지되기 위해서는 그림과 같이 연구에 필요한 빅데이터 수요조사 및 활용도 검토 등을 통해 수요자 등 이용자 중심의 데이터 탐색 및 선정, 대상 빅데이터의 검토 및 서비스 유형 분석을 통해 데이터의 수집 및 저장을 체계화할 수 있는 〈그림 3-3〉과 같은 절차가 필요하다.



〈그림 3-3〉 빅데이터 수집-저장 프레임워크

첫째, 수요 빅데이터 탐색 및 선정 단계에서 공공데이터포털의 공공데이터 활용 신청 순위 등 주요 빅데이터 서비스의 이용자 수요를 분석하고 연구자를 대상으로 환경 분야 데이터 활용사례조사 등을 실시하여 연구수요를 반영한다. 이때, 이미 발간된 보고서에 대

한 검토를 수행하는 것도 좋으나 이 경우 특정 인원이나 특정 조직이 전 대상기간에 대해서 검토하는 것이 좋다.

둘째, 대상 빅데이터 검토 및 분석 단계에서 한국환경공단의 대기질 빅데이터처럼 대상이 선정되면, 해당 빅데이터 서비스를 검토한다. 예를 들어, 한국환경공단에서 제공하는 에어코리아 서비스와 공공데이터포털의 한국환경공단 데이터의 데이터 서비스 유형을 분석한다. 에어코리아는 웹서비스를 통해 일정 조건을 검색한 후, 데이터 테이블을 생성하여 웹으로 표시하여 주고 공공데이터포털의 오픈API는 RESTful 방식으로 조건을 통해 데이터를 요청하고 데이터를 제공한다. 데이터 서비스의 분석을 통해 자동화 소프트웨어를 디자인한다. 코드 구현은 다음 단계에서 실행한다.

셋째, Burp Suite[46], HttpWatch[47], Paros[48] 등 웹서비스 분석용 소프트웨어, 웹브라우저에서 제공하는 개발자 도구 및 자바(Java)[49], 파이선(Python)[50] 등 데이터 수집용 소프트웨어 개발을 위한 프로그래밍 언어[51]를 이용하여 데이터 수집을 자동화하는 코드를 작성하고 PostgreSQL[52], MariaDB[53], MySQL[54], MongoDB[55] 등 데이터베이스(DBMS)[56]에 저장한다. 웹브라우저에서 제공하는 개발자 도구를 통해 브라우저에 보이는 웹 구조를 이해하고 웹서비스 분석용 소프트웨어를 통해 데이터 서비스의 데이터 흐름을 이해해야 데이터 수집을 위한 코드를 작성할 수 있다. 또한 데이터베이스(DBMS)에 데이터

---

46) 웹 보안취약점을 점검하기 위한 소프트웨어, https://portswigger.net/burp/freedownload.
47) 웹 오류를 디버깅하기 위한 소프트웨어, https://www.httpwatch.com/download.
48) 웹 보안취약점을 점검하기 위한 소프트웨어, https://sourceforge.net/projects/paros.
49) 썬 마이크로시스템즈(Sun Microsystems)의 제임스 고슬링(James Gosling)이 개발하고 1995년 썬(2010년 오라클이 인수)이 발표한 플랫폼 독립적인 객체지향 프로그래밍 언어, https://go.java.
50) 1989년 귀도 판 로썸(Guido van Rossum)이 창시하고 1991년 발표된 기존 프로그래밍 언어에 비해 쉬워 실 사용률과 생산성이 좋은 프로그래밍 언어, https://www.python.org.
51) 프로그래밍 언어를 선택하는 기준은 특별히 없으나 목표 분석플랫폼을 고려한 선택이 중요하다고 판단된다.
52) SQL 표준 지원 및 기능이 우수하고 PostGIS를 통한 공간정보(Geospatial query) 활용이 우수한 오픈소스 관계형 DBMS, https://www.postgresql.org.
53) MySQL이 오라클로 인수된 뒤 MySQL AB 출신 개발자들이 개발하고 있는 MySQL과의 호환성 및 성능을 향상시킨 오픈소스 관계형 DBMS, https://mariadb.com.
54) MySQL AB(썬 마이크로시스템즈에 인수 후, 오라클로 다시 인수됨)가 만든 관계형 DBMS로 상용 및 무료 버전을 제공하며 상용 버전은 소스코드를 수정해도 공개하지 않아도 됨, https://www.mysql.com.
55) JSON 형태의 동적 스키마형 문서를 사용하는 NoSQL DBMS, https://www.mongodb.com.
56) 데이터베이스를 선택하는 기준은 데이터 및 서비스 유형을 고려하여 판단하는 것이 중요하다.

를 저장하기 위해서는 앞에서 언급한 데이터 명세서, 데이터 수집 요구사항 명세서, 데이터 테이블 명세서, ERD를 작성하는 것이 좋다.

마지막으로 위와 같은 빅데이터 수집-저장 프레임워크가 원활히 운영될 수 있는 환경을 만들어야 한다.

1) 조직 외부에 Github 등 공개된 온라인 공간에 데이터 소스를 공개하여 미반영 데이터 소스에 대한 요구 및 기반영 데이터 소스의 오류를 수정할 수 있도록 한다.
2) 조직 내부에 빅데이터 활용을 위한 메타데이터 명세서와 빅데이터 활용 가이드를 제공하여 R과 파이선(Python) 등, 연구자 등의 이용자가 데이터를 직접 접근하여 분석 및 가공할 수 있도록 한다. 가급적이면 연구과정에서 생산되는 데이터를 재생할 수 있는 데이터 아카이브와 같은 체계를 도입 및 구축하는 것이 중요하다.
3) 또한 데이터 변경통제위원회와 같은 데이터 추가-수정-삭제-검토 등을 수행할 수 있는 협의체계를 만들어 데이터 활용체계의 선순환 고리가 안착되도록 대응할 필요가 있다.

빅데이터는 짧은 시간에 한 번에 만들어질 수 있는 것이 아니기 때문에 지속적으로 검토하고 대응할 수 있는 절차와 프레임워크, 대응체계 마련이 필요하다.

제4장

결론 및 제언

1. 결론

다양한 환경 분야의 빅데이터를 활용하여 연구를 수행하기 위해서는 빅데이터에 대한 이해를 바탕으로 활용하고자 하는 빅데이터를 탐색하고 이를 제공하는 빅데이터 서비스를 분석해야 한다. 또한 이러한 과정을 통해 수집-저장된 빅데이터는 해당 연구에만 사용되고 버려지는 것이 아니라 연구자들 간에 공유되어야 할 것이다.

많은 연구자들이 데이터를 찾아서 수집-저장하는 과정을 반복하고 있다. 대국민적 관심이 높아지는 환경 데이터일수록 연구주제로 활용하고자 하는 의지가 높기 때문에 시기의 차이가 있을 뿐 연구자들의 수요가 높은 편이다. 많은 연구자들이 연구기간에 많은 시간을 할애하여 데이터를 수집-저장하지만, 수집방법을 절차화하거나 공유를 위한 관리체계를 마련한 사례를 찾아보기 어렵다. 결국 많은 연구자들의 노력이 일회성에 그쳐, 이를 재사용할 수 있는 방안 마련이 시급하였다.

본 연구에서는 대국민적 관심이 높은 빅데이터와 연구자들의 수요를 분석하여 한국환경공단에서 제공하는 대기질 빅데이터를 검토하고 해당 빅데이터 서비스를 분석하여 이 과정을 사례로 수집방법을 정리하였다. 또한 수집방법에 대한 수집-저장절차를 모색하고 프레임워크(안)으로 제시하여 빅데이터의 재사용성과 절차적인 활용방안을 검토하였다.

도출된 수집-저장절차와 프레임워크(안)이 대기질 빅데이터를 기준으로 마련되었다고 하더라도 다양한 데이터에 적용될 수 있는 방안을 제시하였고 실제로 다양한 빅데이터에 적용되고 업데이트될 수 있도록 깃허브를 통해 공개하고 지속적으로 연구할 계획이다. 이를 통해 수집방법이 안정화될 수 있기를 기대한다.

2. 정책 제언

본 연구를 수행하면서 안타까웠던 점은 제공되는 빅데이터 분야와 양에 비해 제공되는 빅데이터 서비스의 품질이 질적으로는 미흡하다는 점이다. 공공데이터포털에서 제공되는 서비스라고 하더라도 동일한 동작 기전이나 적용 변수, 사용법, 오류 메시지들이 모두 제각각이라는 점에서 불편을 겪었다. 또한 제공되는 데이터가 실시간 정보 제공을 위한 대시보드에 활용 가능한 형태로 제공되어 시간적으로나 공간적으로 그 범위가 넓은 연구에 활용하기에는 데이터와 빅데이터 서비스가 부족하였다. 환경 연구에 해당 데이터를 활용하기 위해서는 목적에 맞게 활용할 수 있는 장치가 마련되었으면 한다. 대국민과 밀접한 연관성이 있는 환경 빅데이터가 적극 활용되어야 환경정책 입안자들에게 현실성 있는 연구수행결과가 전달될 것이고, 이를 통해 대국민적 수요를 만족시킬 수 있는 환경정책을 입안할 수 있으리라 기대한다.

또한 장시간의 (고)해상도, 고품질의 환경 빅데이터는 일부 연구자들만 사용하고 버려져서는 안 된다. 수집-저장절차와 프레임워크(안)를 공유하여 확장하고 다듬어서 장기적으로 유지 관리할 수 있는 관리체계를 마련하여야 한다. 연구자의 빅데이터 수요도 주기적으로 반영하고 데이터 수집-저장-활용 로드맵이 마련되어 환경 빅데이터에 대한 수요예측 및 데이터 발굴로 이어져 적시에 연구에 활용될 수 있도록 정비하여야 한다. 시의성 있는 환경연구 결과를 도출하기 위해서는 해당 환경 연구에 활용 가능한 환경 빅데이터의 수집-저장이 선행되어야 하므로 대상 데이터를 활용하는 연구기관에서는 조직 차원에서 관리하였으면 한다.

| 참고문헌 |

[국내문헌]

미래환경대응 정보화 전략 TF(2016), 「미래환경대응 정보화 전략 TF 결과보고서」, p.12, p.58.

Richard L.(2015), 「하파이썬 웹 스크래핑」, 김영하 역, 에이콘출판주식회사

Tom W,(2009) 「Hadoop The Definitive Guide」, 장형석, 장정호, 임상배, 김훈동 역, 4판, 한빛미디어, p.39, p.40.

[온라인 자료]

가트너 블로그, "3D Data Management Controlling Data Volume, Velocity, and Variety", https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, 검색일: 2017.5.1.

DELL EMC, "한국EMC, 대한민국 디지털 데이터의 양 조사 2014년 1,360억 기가바이트(GB) 생성, 2020년에 약 8,470억 기가바이트(GB)에 달할 것", https://korea.emc.com/about/news/press/2014/20140612.htm, 검색일: 2017.7.31.

법제처, "공공데이터의 제공 및 이용 활성화에 관한 법률", http://law.go.kr, 검색일: 2017.7.31.

CERN dashboard, http://dashboard.cern.ch, 검색일: 2017.8.14.

IEA 통계포털, https://www.iea.org/statistics/, 검색일: 2017.9.13.

IEA 데이터포털, http://data.iea.org, 검색일: 2017.9.13.

IIASA Models, Tools, and Data, http://www.iiasa.ac.at/web/home/research/modelsData/models-tools-data.html, 검색일: 2017.9.13.

World Bank Open Data, https://data.worldbank.org, 검색일: 2017.9.13.

OECD Data, https://data.oecd.org, 검색일: 2017.9.13.

U.S. Geological Survey Science Data Catalog, https://data.usgs.gov, 검색일: 2017.9.13.
공공데이터포털, https://www.data.go.kr, 검색일: 2017.10.10.
에어코리아, http://www.airkorea.or.kr, 검색일: 2017.10.10.

부 록

부록 Ⅰ. 환경 데이터 활용사례 설문지(예시)

1. 내가 활용하는 데이터는 다음과 같다.

   □ 물 □ 대기 □ 온실가스 □ 기후변화 □ 자연/생태계 □ 폐기물 □ 신재생 에너지 □ 환경관리
   □ 환경법규 □ 환경보건 □ 환경평가 □ 실내환경 □ 화학물 □ 교통환경 □ 소음/진동 □ 신기술
   □ 환경통계 □ 기타 ( )

2. 내가 사용하는 데이터는 계측치 또는 확정치이다.

   □ 계측치(실시간 데이터) □ 확정치(확정 데이터) □ 기타 ( )

3. 내가 활용하는 데이터는 어디에 있다.

   □ KEI □ 국내 □ 해외 □ 어디에도 없다. 이유는? ( )

4. 내가 활용하는 원외 데이터는 이곳에서 얻는다.

   □ 공공데이터포털 □ 에어코리아(한국환경공단) □ 기상자료개방포털, 기후정보포털 등(기상청)
   □ 국가수자원관리종합정보시스템(국토교통부) □ 물환경정보시스템(국립환경과학원)
   □ K-water 공공데이터개방포털 □ 국가상수도정보시스템(한국환경공단)
   □ 실시간 수질정보시스템(국립환경과학원) □ 토양지하수정보시스템(국립환경과학원)
   □ 전국오염원 조사 온라인 시스템(국립환경과학원)
   □ 국가 대기오염물질 배출량 서비스(국립환경과학원) □ 자원순환정보시스템(한국환경공단)
   □ 환경통계포털(환경부) □ 국가통계포털 등 통계청 □ 화학물질정보시스템(국립환경과학원)
   □ 국가교통DB(한국교통연구원) □ 기타 ( )

5. 내가 활용하는 데이터 관련 소프트웨어는 다음과 같다.

   □ 엑셀 □ R □ 파이선(Python) □ C++ □ FORTRAN □ SQL 소프트웨어 □ SAS □ SPSS
   □ Stata □ SigmaPlot □ GAMS □ IDL □ Matlab □ Mathemetica □ ArcGIS □ QGIS
   □ AutoCAD □ ENVI □ Tableau □ Elastic Stack(Kibana 등) □ 기타 ( )

6. 데이터 기반 연구(또는 업무) 시 불편한 점은?

   □ 가져다 쓰고 싶지만 필요한 데이터 없다. □ 데이터가 있다는데 어디서 받는지 모르겠다.
   □ 데이터 신청절차가 불편하다. □ 데이터가 있는데 받기 어렵다.
   □ 데이터를 받았는데 전처리가(사용하기가) 불편하다.
   □ 데이터 관련 소프트웨어를 사용할 줄 모른다. □ 기타 ( )

7. 데이터 기반 연구(또는 업무) 시 필요한 것은? (서술식)

작성자 : 부서 [ ] 이름 [ ]

부록 II. NoSQL 데이터베이스 종류[57)]

데이터베이스(DBMS) 구분	데이터베이스(DBMS)
Wide Column Store / Column Families	Hadoop / HBase, MapR, Hortonworks, Cloudera, Cassandra, Scylla, Hypertable, Accumulo, Amazon SimpleDB, Cloudata, MonetDB, HPCC, Apache Flink, IBM Informix, Splice Machine, eXtremeDB Financial Edition, ConcourseDB, Druid, KUDU, Elassandra
Document Store	Elastic, ArangoDB, OrientDB, gunDB, MongoDB, Cloud Datastore, Azure DocumentDB, RethinkDB, Couchbase Server, CouchDB, ToroDB, SequoiaDB, NosDB, RavenDB, MarkLogic Server, Clusterpoint Server, JSON ODM, NeDB, Terrastore, AmisaDB, JasDB, RaptorDB, djondb, EJDB, densodb, SisoDB, SDB, NoSQL embedded db, ThruDB, iBoxDB, BergDB, ReasonDB, IBM Cloudant, BagriDB
Key Value / Tuple Store	DynamoDB, Azure Table Storage, Riak, Redis, Aerospike, LevelDB, RocksDB, Berkeley DB, Oracle NOSQL Database, GenieDB, BangDB, Chordless, Scalaris, Tokyo Cabinet / Tyrant, Scallen, Voldemort, Dynomite, KAI, MemcacheDB, Faircom C-Tree, LSM, KitaroDB, upscaledb, STSdb, Tarantool/Box, Chronicle Map, Maxtable, quasardb, Pincaster, RaptorDB, TIBCO Active Spaces, allegro-C, nessDB, HyperDex, SharedHashFile, Symas LMDB, Sophia, NCache, TayzGrid, PickleDB, Mnesia, LightCloud, Hibari, OpenLDAP, Genomu, BinaryRage, Elliptics, DBreeze, TreodeDB, BoltDB, Serenety, Cachelot, filejson, InfinityDB, SCR Siemens Common Repository
Graph Databases	Neo4J, ArangoDB, OrientDB, gunDB, Infinite Graph, Sparksee, TITAN, InfoGrid, HyperGraphDB, GraphBase, Trinity, AllegroGraph, BrightstarDB, Bigdata, Meronymy, WhiteDB, Onyx Database, OpenLink Virtuoso, VertexDB, FlockDB, weaver, Execom IOG, Fallen 8
XML Databases	EMC Documentum xDB, eXist, Sedna, BaseX, Qizx, Berkeley DB XML, JEntigrator
Multimodel Databases, Object Databeses, Grid & Cloud Database Solutions, Multidimensional Databases, Multivalue Databases, Event Sourcing, Time Series / Streaming Databases 등 다양한 성격의 NoSQL이 존재함	

57) NoSQL 데이터베이스 종류는 NoSQL(http://nosql-database.org)에 게시되어 있는 「LIST OF NOSQL Databases」를 기준으로 도표화하여 작성함

Abstract

A Study on Environmental Big-data Scraping Method: Focused on Air Quality Data

Benjamin KJ Han

The purpose of this study is identify the big data that can be used for environmental research through understanding the big data which is the basis of intelligent information society and to develop a procedure and framework of environment big data. In order to using the big data as a center of future and research paradigm, it is necessary to understand and actively apply the big data. In addition, identification and countermeasures for environmental data should be prepared. As a case study, it analyzed the air quality data and services of Airkorea, the process of scraping and storing the big data through service analytic process and presented a framework for scraping method.

Keywords : Big data, Scraping, Storing, Web crawling, Framework

저자약력

한국진 (연구책임)

서강대학교 소프트웨어공학 석사
한국환경정책·평가연구원 선임전문원(현)
E-mail : kjhan@kei.re.kr

주요 연구실적

- 기후변화 적응 정보 공동활용 체계 강화 (2016)
- 2013년 환경영향평가 정보지원시스템 개선 및 운영 (2013)

강성원

미국 루트거트대학교 경제학과 이학 박사
한국환경정책·평가연구원 선임연구위원(현)
E-mail : swkang@kei.re.kr

김도연

충북대학교 정보보호경영학 석사
한국환경정책·평가연구원 연구원(현)
E-mail : dykim@kei.re.kr

김영인

창원대학교 전자계산학 석사
한국환경정책·평가연구원 선임전문원(현)
E-mail : yikim@kei.re.kr

| KEI Working Paper 목록 | 2015~2017

2017년 2017-01 불확실성과 학습효과를 반영한 기후경제 모형 방법론 연구(황인창)
2017-02 환경경제 분석에서 행위자 기반 모형의 활용 방안 연구(채여라, 정예민)
2017-03 인도 물관리 정책의 비교분석과 환경협력 확대 방향(김익재)
2017-04 산림경영사업지의 개발용지 전환 사례조사 및 개선사항의 도출(방상원)
2017-05 환경분야 빅데이터 수집방법 연구(한국진)
2017-06 에머지 방법론을 활용한 유역의 지속가능성 평가: 금강유역을 중심으로(이승준)
2017-07 도시재생 활성화지역 노후 건물의 재정비 시나리오별 환경적 지속가능성 평가를 위한 기초연구(송지윤)
2017-08 Smart waste 및 환경정보 제공을 위한 주민참여형 애플리케이션 활용 연구 (이소라, 임혜숙)
2017-09 2차생성 미세먼지 저감을 위한 암모니아 관리정책마련 기초연구(신동원)
2017-10 주요국가 환경정책 트렌드 분석연구(명수정, 문현주, 신용승, 전호철)
2017-11 한국의 녹색경제지수 산정(김종호)
2017-12 합성생물 관리방안 마련을 위한 국내외 연구동향(오일찬)

2016년 2016-01 시스템과 네트워크 이론을 활용한 미래 환경정책 방향 연구(이승준)
2016-02 공공자료 분석을 통한 친환경적 풍력에너지 개발 기초 연구(김태윤)
2016-03 환경영향평가에서 활용 가능한 주민참여 방법 기초 연구(이상윤)
2016-04 자율주행 자동차의 친환경성 제고를 위한 기초 연구(이승민)
2016-05 미래 고온환경 변화와 직종 간 임금격차 추정(김동현)
2016-06 드론을 이용한 환경재난 사후대응 기술 및 연구동향 분석 연구(손승우)
2016-07 건물부문의 환경 부하 평가 모형 개발을 위한 기초연구(송지윤)
2016-08 근지표환경 임계영역(critical zones)의 환경적 중요성과 환경관리의 미래 이슈(현윤정)
2016-09 시민과학의 자연환경조사 적용방안 연구(김윤정)
2016-10 환경평가 자료의 공공서비스 지원을 위한 기초연구(김태형)
2016-11 토지환경분야의 지속가능발전목표(SDGs) 이행을 위한 정책방향 설정(명수정)
2016-12 건강영향평가 분야에서의 위해소통을 위한 리스크 테이블 제작 연구(하종식)
2016-13 해외 환경정책 인벤토리 구축 연구: 환경전략/대기환경/물환경/국토자연/자원순환 부문 (조일현, 공성용, 한대호, 홍현정, 한상운)
2016-14 해외 환경정책 인벤토리 구축 연구: 환경평가 부문(박하늘)
2016-15 해외 환경정책 인벤토리 구축 연구: 온실가스 감축 부문(김이진, 간순영)
2016-16 지하수 개발사업의 환경영향평가 개선을 위한 기초연구(김경호)
2016-17 토양자원 관리를 위한 전략환경영향평가 개선을 위한 기초연구: 도시개발사업을 중심으로(양경)

2016-18 미세조류 바이오매스의 자원화 활용에 대한 연구: 바이오 (기능성)소재를 중심으로 (지민규)
2016-19 2016 국민환경의식조사 연구(곽소윤)

2015년 2015-01 싱크홀 방지를 위한 환경영향평가 개선방안 연구(김윤승)
2015-02 이슈스캐닝(Horizon Scanning) 기법 활용을 통한 물환경관리 부문 이머징 이슈 발굴 연구(한혜진)
2015-03 기후경제통합-지역평가모형(Regional Integrated Assessment Model of Climate and the Economy) 비교분석 및 국내 모형개발을 위한 기초연구(황인창)
2015-04 기후변화로 인한 고온환경 근로자의 작업역량 저하 추정과 공간적 군집 파악(김동현)
2015-05 환경영향평가 설명회·공청회 운영현황 분석(조공장)
2015-06 도로 및 철도 사업의 토양분야 환경영향평가 사례 연구(신경희)
2015-07 빅데이터를 활용한 환경보건서비스에 관한 기초연구(간순영, 윤성지)
2015-08 자원순환분야 지속가능발전목표(SDGs) 이행 기반 마련을 위한 기초연구(임혜숙)
2015-09 내륙습지에 대한 환경영향평가 개선방안 연구 I: 환경부 전국내륙습지 조사 지침(2011)의 적용을 중심으로(방상원)
2015-10 자원순환성 평가제도 대상 확대를 위한 기초연구(이소라)
2015-11 환경소음 빅데이터의 정책 활용성 제고 방안(박영민)
2015-12 인과지도(Causal Loop)를 활용, 미래 물수급관리 정책 지원을 위한 기초연구 (류재나)
2015-13 생물안전 법제 기초연구(홍현정)
2015-14 지방자치단체 환경영향평가 조례 운영현황 및 효율화 방안(선효성)
2015-15 개발사업의 비점오염 영향평가방법 개발을 위한 기초연구(이진희)
2015-16 환경영향평가제도에서의 생태계보전협력금 활용 개선방안(이상범)
2015-17 환경가치 중장기 연구수요 조사(곽소윤)
2015-18 세종특별자치시의 대기질 관리 기획 연구(심창섭)
2015-19 2015 국민환경의식조사 연구(곽소윤)

※ KEI 설립 이후 현재까지의 보고서 원문은 KEI 홈페이지(www.kei.re.kr)에서 보실 수 있습니다.

환경 분야 빅데이터 수집방법 연구

대기질 데이터를 중심으로

KEI 한국환경정책·평가연구원
Korea Environment Institute

30147 세종특별자치시 시청대로 370
세종국책연구단지 B동(과학 · 인프라동) 8~11층
http://www.kei.re.kr

93530
9 791159 801273
ISBN 979-11-5980-127-3

본 책자는 친환경용지로 인쇄되었습니다