과학데이터 보존 및 활용모델에 관한 연구

A Study on a Model for Using and Preserving Scientific Data

김선태(Kim, Sun-Tae)*, 한선화(Hahn, Sun-Hwa)**, 이태영(Lee, Tae-young)***, 김 용(Kim, Yong)****

【초 록】

본 연구의 목적은 기록물로서 과학데이터의 보존 및 활용 모델을 제시하는데 있다. 과학데이터와 관련하여 미국과 영국, 호주, 유럽연합의 국가별 동향을 조사하였으며, 해외 선진 사례를 프로그램 단위로 조사하였다. 조사된 프로그램은 DataCite, WDS, PANGAEA, Dataverse, BSRN, DLESE, GCMD, SEDIS이다. 각각의 프로그램에서 공통된 시사점을 도출하였으며, 이를 기반으로 본 연구에서는 과학데이터 보존 및 활용을 위한 모델을 제시하였다.

【키워드】

과학데이터, 데이터 관리, 데이터 공유, 데이터센터, 연구데이터, 보존

【Abstract】

This study is to suggest a model for preservation and circulation of scientific data as Records. Analysis for national trends in U.S., British, Australia, Europe about scientific data was performed. Foreign advanced programs on management of scientific data were surveyed and analyzed. The analyzed programs were DataCite, WDS, PANGAEA, Dataverse, BSRN, DLESE, GCMD and SEDIS. Common implications were deducted from each program. With the results of analyzing the programs, this study proposed a model for preservation and circulation of scientific data.

【Keywords】

scientific data, data management, data sharing, data center, research data, preservation

1. 서 론

1.1 연구 배경 및 필요성

다양한 디지털 저작도구의 발전과 인터넷의 발전에 따른 정보유통채널의 다변화는 정보의 생산과 이용에 있어서 획기적인 변화를 가져오고 있다. 즉, 과거의 산업사회 기반의 아날로그 환경에서 정보사회 기반의 디지털 환경으로 정보의 패러다임이 변화한다는 것이다. 특히, 정보환경의 변화에 따라 일상의 모든 활동이 정보시스템을 기반으로 수행되고 이를 통하여 수많은 관련 데이터들이 생산되고 데이터베이스에 저장된다는 것이다. 이와 같은 변화에 따라 많은 연구자들은 자신들의 연구와 이에 대한 결과물의 산출에 있어서 보다 용이한 도구를 활용하여 결과물의 유통을 활성화 할 수 있다. 이와 같은 연구 결과물에 대한 생산과 유통구조의 혁신은 자연과학, 사회과학, 인문과학 등의 다양한 학문분야의 연구의 활성화 및 발전을 가져올 수 있는 중요한 기반을 제공하고 있다. 전통적으로 학자들의 수행하는 연구에 있어서의 패러다임은 연구에 활용하는 중심 도구의 유형에 따라 크게 4가지로 분류를 할 수 있다. 첫 번째, 수천 년부터 행해진 방법으로서 관측이나 실험을 통해 연구데이터를 수집 및 생산한 후 이를 기반으로 연구를 진행하는 것을

* 한국과학기술정보연구원 정책연구실 선임연구원(stkim@kisti.re.kr) (제1저자)
** 한국과학기술정보연구원 정책연구실 책임연구원(shhahn@kisti.re.kr) (공동저자)
*** 전북대학교 문헌정보학과 교수(taehyun@jbnu.ac.kr) (공동저자)
**** 전북대학교 문헌정보학과 조교수, 독서문화연구소 연구원(yk9118@jbnu.ac.kr) (교신저자)
논문접수일자 : 2010년 10월 20일 논문심사일자 : 2010년 11월 30일 게재확정일자 : 2010년 12월 10일

들 수 있다. 두 번째, 지난 수백 년 전부터 이론분야에서 출현한 모델링 방법을 통해 연구를 진행하는 방법을 들 수 있다. 세 번째, 지난 십수 년 동안 진행되고 있는 방법으로 컴퓨팅 자원을 활용한 복잡한 현상을 연구하는 방법이 있으며, 마지막으로 최근에 출현한 방법으로써 데이터를 중심으로 한 연구방법이 있다.[1)]

네 번째 연구 패러다임의 중심은 데이터로서 학술과정에서 생산되는 연구데이터 또는 과학데이터를 의미한다. 본 연구에서는 과학데이터라는 용어로서 통칭한다. 일반적으로 과학데이터는 관측, 감시, 조사, 실험, 연구 분석, 계산 등 일련의 과학기술 활동의 결과로서 수치, 공간, 도표, 문서 등의 형태를 취한다. 따라서 그 형식이 복잡하고 널리 배포되며 범위가 넓은 것이 큰 특징이다.[2)]

과학데이터에 대한 학술적 및 경제적 가치에 대한 관심은 최근에 매우 높아지고 있다. 과거에서 현재까지 많은 다양한 학문분야에서 수많은 연구들이 수행되었으며 이를 통하여 수많은 연구 결과물이 생산되고 활용되고 있는 과정에서 결과물들은 학문발전의 중요한 기반요소를 제공하였다. 그러나 연구의 최종 결과물로서 논문 또는 보고서 등에 발표되거나 인용되는 결과물에 대한 관심은 높았으나 결과물을 산출하는 과정에서 활용되는 초기 결과물(Raw Data)들에 대한 연구자들의 관심은 높지 않았다. 이와 같은 초기 결과물은 주로 최종 결과물을 생산 또는 조합하기 위한 산출물로서 연구자들에게 있어서 중요성 있는 데이터로서 인식되지 못하였다. 그러나 최근 연구의 최종 결과물을 포함하여 연구초기에 생산된 모든 결과물에 대한 관심이 급속도로 높아지고 있다. 이와 같은 흐름은 미국, 유업 등을 포함하여 중국 등에서 활발히 진행되고 있다. 많은 국가에서 정책적으로 과학데이터에 대한 활용과 보존에 적극적인 이유는 크게 두 가지 관점에서 논할 수 있다. 첫째는 수많은 연구를 통하여 발표되고 있는 연구결과에 대한 검증(Verification)에 있다고 할 수 있다. 대부분의 발표되고 있는 연구들은 다양한 실험과 경험적 접근을 통하여 수행되며 또한 많은 시행착오를 통하여 연구의 증거자료로서 과학데이터를 생산하고 이를 기반으로 연구자들의 연구에 대한 타당성을 증명하고 있다. 그러나 연구자들이 발표하는 최종 결과물에서 제시하고 있는 연구결과물의 검증에 대한 어려움이 있다는 것이다. 즉, 최종 결과물은 다양한 시도와 반복적인 실험을 통하여 생산되는데 이와 같은 결과물에 대한 검증을 위해서는 연구자가 실험하였던 환경 및 변수를 동일하게 설정하여야 하지만 이와 같은 검증방법은 매우 어렵다는 것이다. 이러한 어려움으로 인하여 최근 우리 사회에서는 연구자들이 발표한 연구결과물에 대한 오류와 연구윤리의식의 결여에 따라 부정확한 데이터의 오사용에 따른 사회적 파장이 빈번하게 발표되고 있다. 따라서 연구 초기부터 생산된 데이터들을 수집하고 이를 보존함으로써 연구자들의 연구 결과물에 대한 검증도구로서 과학데이터를 활용하고자 하는 것이다. 둘째는 연구를 통하여 생산된 데이터의 재사용성(Reusability)에 있다. 자연과학 또는 공학 분야에 있어서 생산되는 데이터는 연구자들의 많은 시간, 노력, 인력 또한 수많은 예산이 투입되는 경우가 빈번하다. 따라서 국가적 차원에 있어서 이처럼 많은 비용과 노력 그리고 시간을 통하여 생산되는 데이터를 해당 연구에서만 활용하는 것이 아니라 이와 유사한 또는 관련성이 높은 연구에 직접적으로 적용되거나 활용할 수 있는 사례가 매우 많다. 따라서 국가적 차원에 있어서 국가 예산 및 자원의 효율적인 활용과 재사용성에 있어서 연구를 통하여 생산된 모든 데이터 즉, 과학데이터는 매우 중요한 국가 자원이라고 할 수 있으며 이와 같은 자원의 수집, 저장 및 보존은 매우 중요한 업무라고 할 수 있다. 이와 같은 데이터를 중심으로 하고 있는 연구분야의 흐름에 따라 과학데이터에 대한 보존과 활용에 대한 연구가 매우 요구된다.

1.2 연구 목적 및 내용

과학데이터는 연구를 수행하는 과정에서 생산되는 모든 데이터를 포함하며 이러한 데이터는 통계수치, 수식, 이미지지도, 도표, 문서 등을 포함한다. Koski(2009)가 조사한 바에 따르면 생명정보 분야의 경우, 영국 생명정보 데이터베이스가 1년도 되지 않아 3배의 규모로 성장하였으며, 환경과학 분야의 경우, 미국 NASA의 '지구 관측 시스템 데이터와 정보시스템'의 데이터는 2000년부터 2005년까지 데이터양에 비해 2009년 현재 10배 성장하였

1) The FOURTH PARADIGM Data-Intensive Scientific Discovery. Tony Hey.
2) Jinpei Cheng, THE DEVELOPMENT OF CHINA'S SCIENTIFIC DATA SHARING POLICY, Strategies for Preservation of and Open Access to Scientific Data in China: Summerary of a Workshop, 〈http://www.nap.edu/catalog/11710.html〉.

다. 소립자 물리학 분야의 경우, 제네바에 있는 CERN[3]의 LHC[4]는 매년 16Petabytes[5] 정도의 데이터를 생산하고 있다. 정보의 생산 및 유통과 관련된 정보기술의 발전은 과거와는 달리 정보의 생산에 있어서 기하급수적인 증가를 초래하고 있다. 따라서 정보가 늘어나는 것이 문제가 아니라 원하는 정보를 찾기 어려운 것이 문제인 것처럼 과학데이터의 양은 폭발적으로 증가하고 있으며 증가하는 데이터를 검색하고 활용하는데 따른 문제점을 방지하고 구축된 과학데이터의 재사용을 통한 연구의 경제성을 제고하기 위해서도 체계적인 관리가 필수적이다. 이와 같은 '데이터 홍수' 현상의 또 다른 단면으로서 생산된 데이터가 체계적으로 축적되거나 관리되지 못해 '데이터의 유실'이 발생되는 현상이 존재하고 있다. 이와 같은 데이터 유실의 대표적인 사례로서 학술적 연구를 통하여 생산되는 기록물 즉, 기초데이터로서 과학데이터(Scientific Data)가 대표적이라고 할 수 있다. 현재 대부분의 과학데이터는 컴퓨터 소프트웨어나 첨단장비를 통하여 디지털 형태로 생산된다. 그러나 아직까지 과학데이터의 공유나 유통의 중요성에 대한 연구자들의 인식부족과 과학데이터의 수집 및 관리를 지원할 수 있는 국가차원의 관리 체계가 구축되지 못하고 있다. 이와 같은 현실적 문제로 인하여 생산된 과학데이터의 위치정보 파악의 어려움과 함께, 생산된 과학데이터는 해당 데이터의 생산과 관련된 일부 연구자들에 의해서 사용되다가 사장되는 경우가 대부분이라고 할 수 있다. 따라서 본 연구에서는 연구의 부산물로서 생산되는 과학데이터를 과학분야의 중요한 기록물(Records)로서 연구자들에 의해 수행된 연구결과물에 대한 검증과 경제적 관점에서 수많은 예산이 투입되어 생산되는 결과물의 공유와 재활용을 위한 방법을 제안하고자 한다. 이를 위하여 본 연구에서는 과학데이터에 대한 기본 개념 및 특성을 분석하기 위하여 문헌조사 및 학자들에 대한 인터뷰를 수행하였다. 특히 국가차원의 과학데이터 관리 모델을 수립하기 위하여 과학데이터 수집 및 관리 정책을 수립하고 이를 활용하고 있는 국가들의 과학데이터 관련 정책과 모델에 대한 사례조사 및 분석을 수행하였다. 이와 같은 조사 및 분석을 통하여 국내 정보관리 및 유통과 관련된 기관 및 서비스 환경을 반영한 과학데이터 관리모델을 제시하였다.

이를 보다 구체적으로 알아보기 위하여 국가적 차원에서 과학데이터의 수집 및 관리를 수행하고 있는 미국, 호주, 영국 등의 보존과 활용사례를 조사하였다. 특히, 각각의 사례별 특징과 관리방법에 대한 특이 사항을 도출하였으며 이를 기반으로 기록물로서 가치가 있는 과학데이터의 보존과 활용을 위한 모델들을 조사 분석하였다. 세부적으로는 국가차원의 동향과 선진 우수 프로그램을 나누어 조사하였다. 국가차원의 동향은 미국, 호주, 영국, 유럽연합을 조사하였으며, 선진 우수 프로그램은 DataCite외 5개의 프로그램을 조사하였다. 이와 같은 사례조사 결과를 기반으로 국가적으로 요구되는 과학데이터 보존 및 활용을 위한 모델을 제안하였다.

1.3 선행연구

과학데이터는 과학연구활동의 부산물로서 생산되는 기초데이터이다. 국내에서도 정보관리 및 유통의 관점에서 수집 및 관리와 관련된 연구가 일부 진행 되었다. 국내에서는 채균식과 이응봉(2003)은 연구과정에서 생산된 참조과학데이터로서 정의하고 이를 관리하기 위한 국가 참조표준데이터로서 관리의 필요성에 대하여 언급하면서 국가참조표준센터의 설립 방안을 제시하고 있다. 그러나 과학데이터의 수집 및 관리의 중요성에도 불구하고 국내에는 후속연구가 거의 진행되지 못하였다. 한편, 기록관리의 관점에서 이공계열의 측정 및 수치데이터와 사회과학분야의 통계데이터에 대한 관리의 필요성에 대한 연구가 부분적으로 진행되었다.

한편, 대상을 과학데이터가 아닌 일반 기록데이터 및 학술정보에 대한 디지털 아카이빙에 대한 연구로서 정혜경(2004)은 디지털 아카이빙 전략에 있어서 디지털 아카이빙의 무형적 가치를 반영하여 분석하는 모형의 필요성을 주장하였다. 또한 정보경제학 측면의 가치사슬 개념을 적용하여 종합적인 경제성 분석 모형을 제시하면서 가치가속과 가치연결 개념을 도입하여 디지털 아카이빙의 경제성에 대한 비교분석을 통하여 경제성의 요인들을 제시하였다.

이수상(2004)은 디지털보존의 기본적인 개념을 형성

3) European Organization for Nuclear Research 유럽원자핵공동연구소.
4) Large Hadron Collider 대용량 소립자 충돌형 가속기.
5) 페타바이트(Petabyte, PB)는 1015를 의미하는 SI 접두어인 페타와 컴퓨터 데이터의 표시단위인 바이트가 합쳐진 자료량을 의미하는 단위이다. 1 PB = 1015 bytes = 1,000,000,000,000,000 bytes(출처: 위키백과).

하는 요소로서 디지털객체의 수명주기, 아카이빙 워크플로우를 시스템의 입장에서 정의하면서 기술보존, 에뮬레이션, 파일재생, 포맷전환, 캡슐화 등 디지털보존에 사용되는 주요한 보존처리 기술을 비교분석하였다.

박현영과 남태우(2004)는 디지털 아카이빙 정책에 있어서 디지털 아카이빙 환경과 목표에 따라 규정되어지는 정책에 대한 이론적 배경을 분석하였으며 호주 국립도서관의 디지털 아카이빙 정책과 프랑스 통계자료를 위한 보존 정책을 중심으로 사례 분석을 수행하였다. 또한 국내외 학술논문 아카이빙 데이터의 활용성을 증진시킬 수 있는 시스템 및 서비스 전략을 중심으로 논의하고자 한다.

또한 정영임, 최호남과 최선희(2010)는 과학데이터가 아닌 학술논문에 대한 디지털 아카이빙 데이터의 활용성의 확대를 위하여 KISTI의 전략을 국내외 연구와 비교하였으며 NDA 체제 구축의 비용편익 분석을 통해 아카이빙 데이터 활용성 증진을 위해 정책적, 법제적 기반 마련 방안과 아카이빙 데이터의 고부가가치 서비스 제공 방안을 제안하였다.

2. 과학데이터 개념 및 특징

2.1 과학데이터 개념 및 정의

연구자들이 연구행위 또는 활동의 과정에서 연구의 부산물이자 결과물로서 데이터가 생산된다. 이와 같은 데이터에 대하여 학술적으로 정의하면 과학 데이터(Scientific Data), 연구 데이터(Research Data) 또는 연구과학 데이터(Research Scientific Data) 등으로 다양하게 불리고 있다. 그러나 국외의 경우에 있어서 과학데이터의 수집 및 관리의 중요성 및 필요성에 대하여 일찍부터 논의와 연구가 진행되었다. 최근 2010년 6월 독일 하노버에서 개최된 DataCite Summer Meeting에는 과학데이터와 관련된 다양한 분야의 글로벌 전문가들이 모여 과학데이터에 대한 논의를 진행하였다. 이와 같은 논의의 과정에서 연구결과물로서 데이터에 대한 정의에 있어서 과학데이터와 연구데이터라는 용어에 대하여 명확한 구분은 필요치 않으며 일반적으로 과학데이터는 연구데이터 범주에 포함 된다는 것이 공통된 의견이었다. 따라서 본 연구에서는 보다 세부적인 형식을 취하고 있는 용어로서 과학데이터라는 용어를 사용한다. 일반적으로 과학데이터란 학술적 연구를 수행하는 과정 중에 관찰이나 실험 또는 원격 탐지나 시뮬레이션 등의 일련의 연구행위를 통하여 수집, 관측, 측정 되는 기초 데이터(raw data)를 의미한다. 컴퓨터를 통하여 생성된 실험데이터, 통계데이터, 단백질 구조이미지, 생물의 표본자료, 천문학의 분광관측(spectral survey) 자료 등이 여기에 해당된다고 할 수 있다. 이와 같은 과학데이터는 유형과 형식에 있어서 매우 다양하다고 할 수 있다. 과학데이터에 대한 수집, 저장, 관리 등에 대한 연구는 여전히 초기단계라고 할 수 있다. 따라서 과학데이터에 대한 정의는 학자마다 부분적으로 다르게 정의하고 있다. Cheng(2006)은 과학데이터는 과학기술 활동의 결과로서 관측(Observation), 감시(Monitoring), 조사(Investigation), 실험(Experiment), 연구 분석(Research Analysis), 계산(Computation) 등의 활동을 통해 생성된 데이터라 정의하였다. OECD(2006)의 경우, 데이터는 과학 연구 수행을 위한 주요한 원천으로 사용하는 사실적인 기록(숫자, 문자정보, 이미지 및 소리)으로 정의하였다. CCSDS[6](2002)는 과학데이터는 전달, 해석 및 가공에 적합하도록 형식을 갖춘, 재해석이 가능한 정보의 표현이라 정의하였다. 이와 같이 과학데이터에 대한 다양한 정의를 종합하면 과학데이터란 연구자의 연구 활동 과정 중 생성되는 다양한 유형의 사실적 기록을 의미한다고 정의하고 있다. 이와 같은 과학데이터와 관련된 여러 학자와 기관들의 과학데이터에 대한 정의를 종합하면 과학데이터란 연구자의 연구 활동 과정 중 생성되는 다양한 유형의 사실적 기록을 의미한다. 즉, 연구활동을 통하여 생산된 연구활동의 기록물로서 관측, 감시, 조사, 실험, 분석, 계산 등의 과정을 통하여 생산된 문자, 이미지, 오디오, 동영상 등의 아날로그 및 디지털 형식을 포괄하는 데이터라고 종합하여 정의할 수 있다.

2.2 과학데이터 유형 및 특징

과학데이터의 유형은 연구분야 및 연구방법, 관측장비, 실험장비, 분석방법 등에 따라 다양하다. 이러한 과학데이터는 주로 수치정보, 공간정보, 도표정보, 문서 등의 형태를 띤다. 지구관측 및 환경 분야의 데이터는 주로 관측데이터로서 공간 및 수치정보와 이미지 정보가

6) Consultative committee on Space Data Systems.

주를 이루며, 사회과학분야의 데이터는 주로 설문조사를 통한 통계데이터 형태를 취하고 있다. 컴퓨터과학 분야의 데이터는 주로 도표 또는 수치정보를 취하고 있다. 이처럼 과학데이터는 설문조사를 통해 수집된 소량의 통계데이터 부터 미립자 충돌 가속기를 통해 매년 16 Petabytes씩 생산되는 대용량 데이터까지 규모와 형태적인 측면에서 매우 다양한 특징을 보여주고 있다. 이와 같은 과학데이터가 가지고 있는 특징을 보다 구체적으로 살펴보면 첫째, 데이터의 형식에 있어서 매우 다양하다는 것이다. 위에서 언급되었듯이 과학데이터는 문서형식을 취하는 아날로그 정보형식에서 부터 컴퓨터 파일, 이미지 등의 다양한 디지털 정보 형식을 포괄하여 모든 유형의 형식으로 존재한다. 둘째, 과학데이터는 특정한 현상을 설명하기 위하여 재가공이 가능하다는 것이다. 일반적으로 과학데이터는 최종의 결과물을 추출하기 위한 기초데이터(Raw Data)로서 기능을 수행한다. 따라서 논문, 보고서 등에서 볼 수 있는 연구자들의 연구행위 및 활동에 대한 최종 결과물을 생산하기 위하여 연구과정에서 생산된 다양한 기초데이터를 활용한다. 셋째, 과학데이터는 자연과학, 공학 등의 분야뿐만이 아니라 인문과학, 사회과학 등에서 생산되는 다양한 통계데이터까지를 포괄한다. 따라서 생산되는 학문의 영역이 매우 다원화하다는 것이다. 넷째, 데이터 형식의 다양성으로 인하여 관리의 어려움이 존재한다는 것이다. 즉, 아날로그 및 디지털 형식으로 존재함으로 해서 단순히 데이터베이스에 저장하는 것이 아닌 체계화된 관리방법이 요구된다. 이상에서 살펴본 바와 같이 과학데이터가 가지고 있는 다양한 형식 및 관리적 특징으로 인하여 좁게는 연구자 개인적인 측면에서 또한 넓게는 국가적인 측면에서 수집, 저장 및 관리에 있어서 체계화된 정책과 관리방안이 요구된다. 특히, 국가차원에서 지원되고 있는 다양한 연구활동들을 통하여 생산되는 과학데이터는 국가적 연구경쟁력의 확보와 연구예산이 효율적인 활용이라는 측면에서 깊이 있는 논의와 체계화된 관리방안의 수립이 절실히 요구된다.

3. 국가별 동향

3.1 미국

미국의 연방기구들은 디지털데이터에 대한 관리 및 수집을 위하여 '연방기구 간의 워킹 그룹(Interagency Working Group on Digital Data, IWGDD)'을 조직하였다. IWGDD는 연방기구들에서 생산된 데이터에 대한 공개적인 접근을 가능토록 하기 위하여 목적으로 조직되었다(CUL 2008). 연구비 지원기관인 국립보건원(National Institutes of Health, NIH), 국립과학재단(National Science Foundation, NSF) 등은 연구비 투자 결과의 환원을 최대화 하는데 큰 관심을 갖고 있으며 이를 위한 중요한 방법으로서 해당 기관에서 지원한 연구과제들을 통하여 생산된 데이터에 대한 공유가 하나의 대안이라 판단하였으며 이에 대한 구체적인 실행전략으로서 국립보건원의 경우에 있어서 50만 달러 이상의 연구과제 수행자에게 데이터 공유를 요구하고 있다(NIH 2007).

NASA는 분산된 아카이브 센터를 2007년부터 지원하기 시작하였다(NASA 2007). NSF의 Office of Cyber Infrastructure(OCI)는 총 1억 달러를 투입하여 'Sustainable Digital Data Preservation and Access Network Partners(DataNet)' 프로그램을 2007년에 시작하였는데, 해당 프로그램은 관련 기관에서 생산된 데이터에 대한 공유를 지원하기 위하여 생성된 프로그램이라고 할 수 있다. DataNet 프로그램에 소속된 기관들은 상호운용성(Interoperability)이 가능한 데이터 보존과 접근을 위한 네트워크 노드 역할을 수행한다. 또한 데이터 보존과 접속, 분석, 가시화 등의 기능도 제공한다. OCI는 최근 2010년 7월 30일까지 전산분야의 전문적인 역량을 가지고 있는 프로그램 디렉터를 채용할 계획이며 세부적인 전문분야는 소프트웨어 개발과 데이터 가시화 프로그램이다. 이와 같은 계획은 데이터 보존과 활용을 최대한 지원하고자 하는 노력으로 판단된다. NSF는 DataNet 이외에도 지구과학, 환경관측, 극지연구 등에 있어서 가상공간상의 기반체계(Cyber Infrastructure)와 데이터의 수집과 관리(Data Curation)를 위한 제안을 요청하였다.

3.2 호주

호주의 연구커뮤니티에서 과학데이터에 대한 관리와 관련하여 호주 연방정부로부터 예산을 지원받아 개발 중인 국가 프로젝트로서 ANDS(Australian National Data Service)가 운영 중에 있다. ANDS는 국가정책에 영향을 주고 성공적인 데이터 큐레이션 확산에 힘쓰며 이질적인 연구데이터 컬렉션을 상호 조화롭게 구성된 컬렉션으로 변경하는 것을 목적으로 하고 있다.

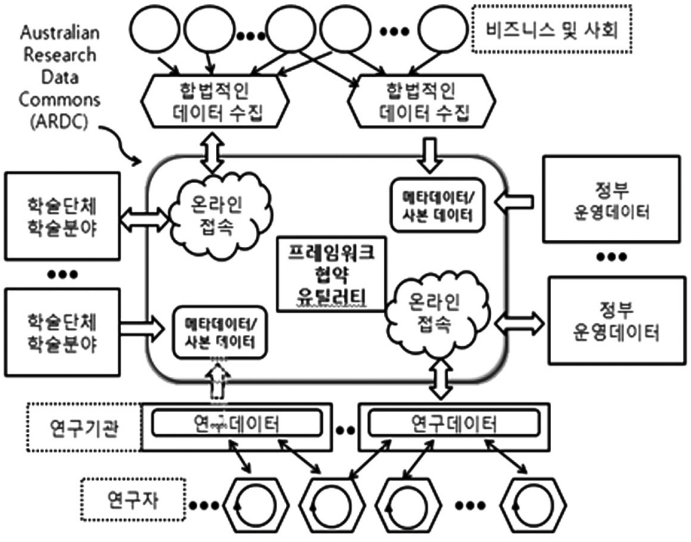
ANDS는 오스트레일리아의 국가협업연구 기반구축 사업인 NCRIS 사업 중 협업 플랫폼 구축(Platforms for Collaboration)프로그램의 일환으로 추진되었다. 이와 같은 ANDS 프로그램이 추진된 배경을 알아보면 다음과 같다. 2006년에 국가적 차원에서의 과학데이터에 대한 관리와 관련한 컨설팅을 통하여 2007년에 대규모 연구포럼과 함께, ANDS 기술워킹그룹(TWG: Technical Working Group)을 구성하였다. ANDS 기술워킹그룹은 과학데이터의 수집, 관리 등과 관련된 다양한 연구보고서를 제출하였으며 이를 기반으로 호주의 DIISR와 EIF의 지원으로 Monash 대학을 중심으로 대학과 국가연구소가 협력하여 2008년 9월에 ANDS를 설립하고 운영하고 있다(박동진 2010). ANDS는 호주의 대학이나 공적자금이 투입되는 연구기관, 정부조직에서 보관중인 연구데이터의 발견과 접근을 지원하기 위해 ARDC(Australian Research Data Commons)[7]를 조직하였다. ARDC는 실제 데이터를 소유하지 않고 단지 데이터에 대한 위치정보와 컬렉션에 대한 기술정보만을 보관하고 있으며 컬렉션 자체는 보관하지 않는다. 아래 〈그림 1〉은 ARDC의 개념도를 나타낸다.

ANDS는 ARCS(Australian Research Collaboration Service)와 유기적인 협력관계를 구축하고 있다. 2007년에 설립된 ARCS는 연구과정을 통하여 생산된 과학데이터를 저장하기 위한 호주의 국가시스템으로써 활용되고 있다. 또한 연구자들의 협업적 연구활동을 지원하기 위해서 데이터베이스 호스팅 서비스를 제공하고 있으며, 연구데이터를 수집하기 위해 연구자가 데이터를 생성하고 기탁하는 환경을 지원하는 서비스까지 제공하고 있다. 2010년 7월 현재, 14개의 프로젝트를 진행하고 있으며 다양한 학문분야 커뮤니티의 신청을 받아 가시화 및 시뮬레이션 개발부터 온톨로지 중심의 데이터 관리에 이르기까지 연구데이터를 관리하고 유통시키기 위한 실질적인 연구를 지원하고 있다. ANDS는 ARCS가 제공하는 데이터 저장소와 ANDS가 운영하고 있는 메타데이터 저장소인 ARDC를 이용해서 서비스 체제를 완성시키고 있다.

3.3 영국

영국의 합동정보시스템위원회(Joint Information Systems Committee, JISC)와 e-Science 핵심프로그램 연구위원회의 협업 프로젝트로서 2004년 설립된 DCC(Digital Curation Centre)는 디지털 큐레이션에 관한 연구를 활발하게 수행하고 있다. 과학데이터를 수집, 관리 및 유통하기 위한 일련의 과정은 데이터 큐레이션과 관련이 있기 때문에 DCC를 사례조사 대상으로 선정하였다.

DCC에서 수행했거나 수행중인 대표적인 프로젝트 5개를 살펴보면 다음과 같다.



〈그림 1〉 ARDC 개념도(박동진 2010)

7) 용어 'Commons'가 의미하는 것은 커뮤니티가 사용할 수 있도록 만들어진 리소스를 의미함.

첫 번째, Incremental project는 2011년 3월에 종료되는 프로젝트로서 직접 연구데이터가 생산되는 곳에 DCC 연구그룹이 파견되어 기존 툴을 사용하도록 하고, 실제 데이터를 구축 및 관리하는데 필요한 일련의 과정들을 연구자와 함께 수행하면서 이론적인 것과 실제의 차이를 극복하기 위한 프로젝트의 성격을 가지고 있다. 두 번째, SCARP는 2009년에 종료된 프로젝트로서 데이터 저장과, 공유, 재사용, 큐레이션, 보존을 실제 연구 프로젝트와 함께 진행한 프로젝트로서 다양한 학문분야의 차이점을 발견하는 것을 목표로 하였다. 세 번째, disciplinary differences는 2009년 8월에 종료된 프로젝트로서 DCC와 에든버러대학의 Institute for the Study of Science, Technology and Innovation(ISSTI) 연구소의 합작 프로젝트로서 생명과학 분야의 7개 연구팀을 대상으로 연구를 수행하였으며 해당 프로젝트는 데이터 큐레이션과 관련하여 학제간 차이점을 분석하는 프로젝트이다. 네 번째, ERIS(Enhancing Repository Infrastructure in Scotland)는 JISC에 의해 예산이 지원되어 2년 동안 수행되는 프로젝트로서 2011년 3월 종료된다. 연구자와 리포지터리 관리자의 긴밀한 협력을 통해 연구자 중심의 기반구조를 구축할 수 있는 솔루션 개발을 목적으로 하고 있다. 마지막으로 Open Science Case Studies는 2009년 11월부터 2010년 2월까지 수행된 짧은 기간의 프로젝트로서 연구데이터와 결과물 공개에 있어서 무엇이 연구자에게 돌아가는 혜택인지를 연구한 프로젝트이다.

3.4 유럽연합

유럽에서는 논문, 보고서, 데이터, 혹은 다양한 유형의 정보가 유럽 전역에서 손쉽게 접근 가능하도록 하기 위해 정보유통 기반구조를 만드는 국제적인 프로젝트로서 DRIVER를 추진 중에 있다. 2006년 6월에 테스트 베드 개발을 시작하여 2007년 11월에 개시된 DRIVER는 2010년 7월 현재 유럽지역 33개국에 있는 249개의 리포지터리를 대상으로 250만 건의 콘텐츠를 대상으로 통합검색 기능을 제공하고 있다. 2010년 현재 유럽 전역에서 운영중인 공개적으로 접근이 가능한 리포지터리는 824개로 전체의 약 30%에 이른다. DRIVER 프로젝트의 목적은 다양한 이용자 그룹에게 텍스트 기반의 콘텐츠뿐만 아니라 과학/기술보고서, 작업문서(working papers), 출판전 자료(pre-prints), 연구데이터 자체(original research data)를 제공하는 것이다.

위의 사례에서 언급하고 있는 사례들은 주로 국가적 차원에서 공공분야에서 수행되고 있는 것으로서 이밖에도 상업적인 기관에서의 과학데이터의 수집에 따른 보존 및 활용과 관련된 사례들도 매우 중요하다. 대표적으로 2008년 1월에 구글은 공개적으로 접근이 가능한 과학데이터에 대한 호스팅 서비스를 계획하고 있음을 선언하였다. 또한 마이크로소프트사는 연구자들의 연구성과물을 관리하는 저장소를 개발 중에 있다(CUL 2008).

4. 과학데이터 관리 프로그램 사례 분석

과학데이터 관련 프로그램은 방대한 분야에서 다양한 목적을 가지고 진행되고 있다. 본 연구에서는 과학데이터의 생산관점 보다는 과학데이터의 활용 및 관리에 따른 유통측면에 초점을 맞추어 사례를 발굴하고 조사하였다. DataCite 외 7개 프로그램의 특징을 집중적으로 조사 하였으며 〈표 1〉에서는 조사 및 분석된 프로그램들에 대한 특징을 기술하고 있으며 해당 프로그램들의 내용을 세부적으로 살펴보면 다음과 같다.

DataCite는 국제적인 컨소시엄으로서 2010년 8월 말 현재 9개국에서 12개 기관이 참여 중이며 DOI(Digital Object Identifier)[8]가 부여된 80만 건 이상의 레코드를 확보하고 있다. DataCite는 과학데이터에 DOI를 부여하는 에이전시 역할을 수행하고 있다.

ICSU의 WDS는 국제과학연맹이사회가 2010년 8월 말 현재 미국, 유럽, 러시아, 일본 등, 12개 국가의 52개 WDC를 2009년에 통합하여 만든 것 이다. WDS는 일곱 개의 목적을 가지고 있다. 첫째, 과학데이터와 정보로의 전 세계적인 공정한 접속을 가능하게 하고 둘째, 데이터로의 접근을 용이하게 하며, 셋째, 데이터로의 손쉬운 접속 보장을 위한 노력을 수행하며 넷째, 품질을 보장하는 데이터와 정보 제공. 다섯째, 개선된 데이터 관리(stewardship)를 촉진하고 여섯째, 정보격차의 해소. 일곱째, 보다 나은 연구(better science)를 위한 데이터 제공이다(WDS 2010).

PANGAEA는 2010년 8월말 현재 170 개의 프로젝트 데이터를 서비스하고 있다. 세계 해양환경 과학 데이터센

8) 논문에 부여되는 글로벌 식별자로 활용되고 있음.

〈표 1〉 과학데이터 관련 주요 프로그램 특징

프로그램	특징
DataCite	- 연구 데이터세트를 적재, 식별, 인용할 수 있도록 지원 - 문헌 이외의 모든 데이터에 대한 기술내용 공개 및 서비스 지원 - 데이터 공개를 위해 필요한 프로세스와 표준을 지원 - 논문의 근거자료가 되는 연구데이터와 논문간의 연계를 제공
WDS	- World Data Services - 국제과학연맹이사회 WDC(World Data Center) 통합 - 독립적으로 존재하던 세계데이터센터들과 개별적인 서비스들 상호 연동 - 새롭게 출현하고 있는 기술들과 과학데이터 관련 결과물들을 포함
PANGAEA	- Open Access 라이브러리로 운영되는 정보 시스템 - 지구시스템 연구로부터 지구참조 데이터 보존, 출판, 배포 목적 - 과학데이터 저장소 및 배포시스템으로 활용 - 과학데이터 해석, 가시화, 탐색을 위해 여러 가지 소프트웨어 무료 제공
Dataverse Network	- 웹상에서 무상으로 공급하고 있는 프로젝트 및 과학데이터 공유 솔루션 - 모든 Dataverse 통합검색을 제공하고 있으며, 지속적 접근을 보장 - 데이터세트 검증과 유효성 체크를 위한 Universal Numerical fingerprint 제공
BSRN	- BSRN(Baseline Surface Radiation Network) - 지구표면 방사선 분야에서 주요한 변화 감지하는 것이 프로젝트의 목적 - BSRN 데이터는 연구를 목적으로 하면 누구나 무료로 사용가능 - 지구과학과 환경과학의 출판네트워크인, PANGAEA를 통한 접근을 제공
DLESE	- 교육자, 학생, 과학자가 참여하는 분산된 커뮤니티 - 자원으로는 전자 자료로서 강의계획서, 지도, 이미지, 데이터 세트, 가시화, 평가도구, 커리큘럼, 온라인 강의 등 다양 - 자원에 대해 기술된 메타데이터만 보유
IODP와 SEDIS	- Scientific Ocean Drilling 관련 모든 데이터와 정보에 대한 접근제공을 목표 - 분산된 데이터를 메타데이터를 이용해 통합함으로써 구현
NASA의 GCMD	- 과학데이터를 효율적으로 관리할 수 있는 채널을 제공 - 온라인에서 중복적인 디렉토리 서비스가 만들어지는 걸 지양 - 극지, 해양 등 총 14개 분야 데이터세트 제공

터(World Data Center for Marine Environmental Sciences, WDC-MARE)는 PANGAEA를 데이터 아카이빙과 배포 시스템으로 사용하고 있으며 Earth System Science Data(ESSD)는 해당 저널의 아카이브로 PANGAEA를 지정하여 서비스하고 있다.

데이터 관리 및 아카이빙 정책은 ICSU WDC 데이터 기준과 OECD 정책을 따르고 있다. 또한 제출되는 모든 데이터는 Creative Commons License를 적용하여 제공된다. PANGAEA는 출판사 사이트와 연계되어 논문과 과학데이터의 연계 서비스를 제공하고 있다(PANGAEA 2010).

Dataverse Network는 하버드대학에서 개발해 웹을 통하여 일반이용자에서 무료로 서비스되고 있다. Dataverse는 여러 개의 Dataverse를 호스팅 함으로써 과학데이터를 공유하며 재활용할 수 있는 솔루션을 제공한다. DOI, Dublin Core, FGDC, MARC 등의 다양한 유형의 데이터 반출포맷을 지원하고 있으며 데이터에 대한 통합검색 프로토콜로서 Z39.50도 지원한다(Dataverse 2010).

BSRN(Baseline Surface Radiation Network)은 세계방사선 감시센터(WRMC: World Radiation Monitoring Center)에서 추진 중인 프로젝트로서 북극과 남극은 물론 중위도, 고위도 해양을 연구하는 Alfred Wegener 연구소에서 운영하고 있다. 기후 변화와 관련이 있는 지구표면의 방사선 분야에서 주요한 변화를 감지하는 것이 해당 프로젝트의 목적이다. WRMC의 모든 데이터는 PANGAEA 안에서 별도의 논리적인 레코드(LR: Logical Records)로 존재한다(BSRN 2010).

DLESE는 미국 국립과학재단의 재정 지원을 받아 개발되었으며, 현재는 국가대기연구센터에서 운영 중에 있다. 연구자가 연구수행 중 생성한 모든 데이터를 연구데이터로 정의할 때 DLESE에서 관리 및 유통되는 자료 중 과학데이터가 상당하다고 볼 수 있다. 또한 DLESE의 지구과학 교육 지원 내용 중 교육에 효율적으로 활용될 수 있도록 여러 도구와 인터페이스를 포함한 데이터 세

트, 이미지 등의 제공은 과학데이터 서비스의 대표적 사례라고 할 수 있다.

IODP(Integrated Ocean Drilling Program)은 웹기반의 정보검색 서비스인 SEDIS(Scientific Earth Drilling Information)를 개발 중에 있다. 과학해양탐사(Scientific Ocean Drilling)과 관련된 모든 데이터와 정보에 대한 접근제공을 목표로 개발 중인 SEIDS는 분산된 데이터를 메타데이터를 이용해 통합하여 개발되고 있다. SEDIS의 주요 데이터 소스로는 IODP implementing organizations (IOs)-United States(USIO)와 일본의 CDEX, 캐나다의 ESO이다. 향후 대륙과 호수 탐사데이터도 포함 할 예정이며, 출판 검색엔진을 포함할 예정이다. 또한 보다 상세한 데이터검색과 가시화 및 매핑 도구를 제공할 계획이다(SEDIS 2010).

지구 변화 마스터 디렉토리(Global Change Mater Directory, GCMD)는 총 14개 분야(농업, 대기, 생물학적 분류, 생물권, 기후지표, 극지(Cryospher), 인체치수(Human Dimensions), 지표, 해양, 극지기온, 지구(solid earth), 스펙트럼/엔지니어링, 태양-지구 상호작용, 지상수권(Terrestrial Hydrosphere))의 데이터 세트를 제공 중에 있다.

5. 과학데이터 보존 및 활용 모델

5.1 모델개발에 따른 고려사항

본 연구에서는 국내에 적용 가능한 과학데이터의 보존 및 활용모델을 제안하기 위하여 과학데이터와 관련된 다양한 해외사례에 대한 조사 및 분석을 통하여 크게 일곱 가지의 시사점을 도출하였다.

첫째, 연구성과물로서 과학데이터에 대한 인식의 확산 및 데이터의 수집, 정리 및 관리와 관련된 국가적 정책수립이다. 연구과정에서 생산되는 과학데이터를 연구성과물로서 인정되고 또한 과학데이터에 대한 중요성에 대한 인식의 확산에 따른 보존 및 공동 활용의 중요한 대상으로서 고려하는 인식이 확산되고 있다. 이와 관련하여 연구비 지원기관을 중심으로 데이터 관련 정책이 만들어지고 있다. 해외 사례에서 살펴본 미국의 NIH 이외에도 11개 연구비 지원기관과 영국의 예술과 인문학(Arts and Humanities) 연구위원회 외 7개 연구비 지원기관 및 호주, 캐나다, 핀란드, 프랑스, 스페인 등의 연구비 지원기관에서도 영국이나 미국과 같이 과학데이터와 관련된 다양한 정책을 수립하고 이를 실행하고 있다. 과학데이터가 손실되는 것을 막아 보존하고 재활용되기 위해서는 이를 제도적으로 뒷받침 해줄 수 있는 정책이 수반되어야 한다.

둘째, 데이터 기술을 위한 이용자 교육이다. 호주의 ANDS와 영국의 DCC 사례를 볼 때 과학데이터 생산 현장에서 부터 데이터를 기술하는 방법을 연구자에게 교육하는 과정은 데이터의 발견, 활용 및 유통에 직접적인 영향을 주므로 매우 중요한 요소 중 하나이다. 발견되지 않는 데이터는 사용되지 않는 데이터라고 할 수 있듯이 데이터에 대한 기술은 데이터의 생명주기에 직접적이고 치명적인 영향을 줄 수 있다. 즉, 데이터 기술에 대한 교육은 데이터를 효과적으로 저장하고 이를 활용하기 위한 기반이 되는 것이라고 할 수 있다

셋째, 과학데이터의 영속적 접근 방법제공이다. DataCite 컨소시엄 및 하버드대학교의 Dataverse, PANGAEA는 과학데이터의 영속적 접근을 제공하기 위하여 DOI 등 다양한 국제적으로 공인된 식별자를 등록하여 서비스에 활용하고 있다. DataCite의 경우는 과학데이터에 글로벌 식별자 DOI를 부여하여 데이터 유통에 새로운 패러다임을 만들고 있다. 이와 같은 시도는 CrossRef와 같이 연구논문에 대하여 DOI를 적용시킨 점과 유사하며 차이점이라고 할 수 있는 것은 DataCite는 과학데이터에 특화되어 있다는 것이다. DOI를 통해 데이터의 물리적인 위치의 변화에 관계없이 데이터에 대한 영속적인 접근을 연구자에게 제공할 수 있게 되었으며 기존 학술정보 유통 채널과의 연계를 통해 연구자에게 영구적인 접근을 허용함으로써 보다 나은 안정적인 연구환경과 정보자원을 제공할 수 있게 되었다.

넷째, 메타데이터에 대한 중앙집중형 기반구조이다. 유럽연합의 DRIVER나 미국의 DLESE, 독일의 DataCite, ICSU의 WDS, PANGAEA, IODP의 SEDIS 사례를 살펴보면, 원시데이터(Raw Data)와 메타데이터를 이원화하여 관리함을 알 수 있다. 원시데이터는 과학데이터를 캡쳐링(Capturing)을 하고 있는 기관에서 보관하고 이를 설명하는 메타데이터만 통합하여 검색 인터페이스를 제공하는 서비스 형태를 취하고 있다.

다섯째, 데이터 아카이빙을 수행하는 데이터센터이다. WDC-MARE의 경우 PANGAEA를 과학데이터의 배포 시스템으로 활용하는 것과 함께, 아카이빙 시스템으로 활용하고 있다. 이는 원격지 데이터 센터와 연계하여 서비스를 제공하는 상위레벨의 서비스를 원격지 재난 방

지 시스템으로 활용하는 사례로서 손쉽게 소멸되는 디지털 자원의 관리를 위한 좋은 협력모델이라 할 수 있다.

여섯째, 데이터전송을 위한 표준프로토콜 사용이다. 호주의 ANDS와 유럽연합의 DRIVER, ICSU의 WDS의 사례를 살펴보면 메타데이터 전송에 있어서 시스템 간의 자동화된 프로세스를 지원하기 위하여 OAI-PMH 프로토콜을 사용하고 있다. 즉, 과학데이터의 자동화된 수집과 배포에 따른 표준화된 프로토콜의 개발은 연구자들의 부주의로 인하여 자칫 유실될 수 있는 상황을 방지하고 안정적인 과학데이터의 수집과 유통을 위한 기반이 될 수 있다.

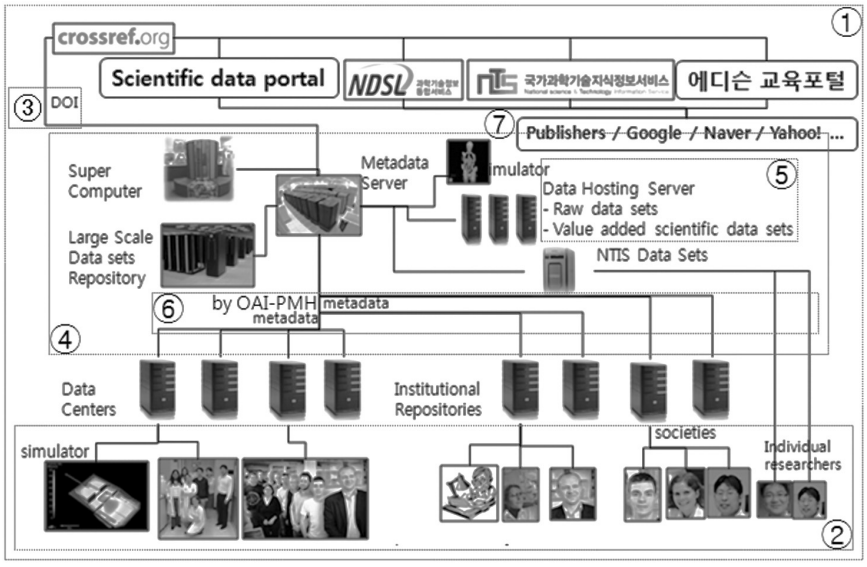
일곱째, 논문서비스와 연계된 과학데이터 서비스이다. PANGAEA의 사례를 보면 엘시비어 출판사의 논문서비스와 PANGAEA의 과학데이터가 연계되어 서비스되는 걸 확인할 수 있다. 물론 그 역도 성립한다. 이는 기존에 존재하던 문헌 위주의 서비스에 새로운 패러다임을 예고하는 사례라 할 수 있다. 이상의 도출된 7가지의 시사점을 정리하면 아래와 같다.

① 연구성과물로서의 과학데이터 인식 확산 및 데이터 관련 정책수립
② 데이터 기술을 위한 이용자 교육
③ 과학데이터의 영속적 접근 방법제공
④ 메타데이터 중앙집중형 기반구조
⑤ 데이터 아카이빙을 수행하는 데이터센터
⑥ 데이터전송을 위한 표준프로토콜 사용
⑦ 논문서비스와 연계된 과학데이터 서비스

5.2 과학데이터 보존 및 활용 모델

과학데이터의 수집 및 활용에 있어서 과학데이터의 생산 및 유통과 관련된 기관 및 연구자들은 생산자이자 소비자로서 역할이 요구된다. 과학데이터의 보존 및 활용모델의 구현을 위해서는 관련 기관, 유통시스템, 유통채널, 관리요소 등에 대한 전반적인 고려가 요구된다. 이와 같은 고려사항을 기반으로 〈그림 2〉는 사례분석을 통하여 도출된 7가지의 시사점을 감안하여 과학데이터를 보존하고 활용하기 위한 모델을 제시하고 있으며 번호는 도출된 7가지의 시사점과 상호 매핑되어 있다. 과학데이터는 체계화된 정보만큼이나 연구자들에게 있어서 매우 가치 있고 중요한 자료라고 할 수 있다. 따라서 과학데이터의 효과적인 활용을 위한 유통 및 수집은 과학데이터 활용의 활성화를 위하여 매우 중요한 요소라고 할 수 있다. 따라서 본 연구에서 제안하고 있는 모델은 과학데이터의 효과적인 수집 및 유통을 위한 단계를 고려하였다. 이를 위하여 국내의 과학기술정보 유통 및 확산에 중요한 기능을 수행하고 있는 KISTI의 과학기술정보 유통 서비스 플랫폼인 NDSL과 NTIS를 포함하고 있으며 향후 추진할 과학데이터 포털 서비스와 에디슨 교육포털 기능도 본 연구에서 제시하고 있는 모델에 포함되었다.

①번의 경우에는 제시된 모델에 나타난 과학데이터의 생산, 소비 및 유통과 관련된 모든 이해관계자들에게 적용될 수 있는 모델과 정책이 요구됨을 표현한다. 따라서



〈그림 2〉 과학데이터 보존 및 활용을 위한 모델

본 연구에서 제안하고 있는 과학데이터 보존 및 활용모델은 단순히 시스템적인 관점이 아닌 과학데이터의 생산, 소비 및 유통의 전 분야를 망라하는 개괄적인 모델이라고 할 수 있다. 즉, 과학데이터의 보존 및 활용에 있어서 과학데이터의 생산자, 소비자, 유통채널, 관리시스템 및 메타데이터를 포함한 관리체계 등의 개별 요소들의 유기적인 협력이 요구된다.

②번은 연구현장에서 데이터를 직접 생산하고 구축하는 연구자들을 위한 교육이 필요함을 나타낸다. 기존의 과학데이터는 최종결과물의 부산물로써 인식되고 있다. 따라서 연구를 수행하고 있는 과학데이터 생산자인 연구자들은 과학데이터의 중요성과 가치에 대한 인식이 부족하였다. 이러한 인식으로 인하여 과학데이터에 대한 관리와 수집이 체계적으로 수행되지 않았으며 관리체계의 구축도 매우 미비하였다. 따라서 국가적 차원에서의 과학데이터의 관리를 위한 관리체계의 구축 및 관리도구의 개발이 요구된다. 연구자들에 대한 과학데이터의 중요성 및 가치성에 대한 교육과 함께, 관리체계 및 관리도구에 대한 교육을 통하여 효과적인 과학데이터의 보존 및 활용이 출발된다.

③번은 과학데이터의 물리적인 위치 변경이 있더라도 데이터로의 영속적인 접근을 허용하기 위해 DOI를 활용하는 것을 나타낸다. 즉, 과학데이터의 관리체계와 도구의 개발을 통하여 과학데이터의 검증 및 재활용을 위한 관리가 확보될 수 있다. 이를 위하여 먼저 개별적으로 생산되는 과학데이터에 대한 영구적인 식별기호를 부여하고 이를 통하여 관리기관 및 관리자에 대한 식별을 통하여 과학데이터가 요구되는 소비자들은 필요로 하는 데이터에 대한 접근(Availability)과 이용성(Usability)을 확보할 수 있다.

④번은 원시데이터를 가지고 있는 데이터 센터 및 기관 리포지터리(Repository)와 메타데이터를 통합적으로 관리하는 센터간의 이중화된 구조를 나타낸다. 과학데이터는 특정 기관 도는 연구자만이 생산하는 것이 아닌 모든 연구기관 및 연구자들이 생산자가 될 수 있으며 각각의 연구기관 및 연구자들은 연구수행과정에서 생산되는 과학데이터를 개별적으로 관리할 수 있다. 따라서 개별적으로 관리되거나 중앙에서 통합 관리되는 과학데이터에 대한 체계화된 관리를 위해서는 메타데이터 관리도구 및 체계를 통하여 통합적으로 관리되어져야 한다. 특히, 중앙의 통합관리기관을 통하여 과학데이터의 공유 및 재활용성을 확보할 수 있다. 이와 같은 체계화된 관리를 통하여 비로소 과학데이터의 보존 및 활용의 가치가 증대될 수 있다.

⑤번은 데이터양은 많으며, 이를 관리할 수 있는 하드웨어적인 인프라나 전문 인력이 부족하여 원시데이터를 대신 관리하는 호스팅 서비스의 모델을 표현한 것이다. 이것은 PANGAEA의 사례와 같이 데이터를 위탁받아 아카이빙 수행하는 모델로도 이해할 수 있다. 연구를 수행하는 기관의 유형, 규모 및 성격에 따라 직접 과학데이터를 관리모델이 달라질 수 있다. 특히, 폭발적으로 증가하는 과학데이터의 양적 수준을 고려한다면 이를 직접 관리하는데 있어서 어려움이 발생할 수 있다. 따라서 과학데이터를 직접 관리하기 어려운 기관 또는 연구자들은 생산된 과학데이터를 전문화된 관리기관에 이관함으로써 관리에 따른 어려움을 해소할 수 있다. 전문관리기관은 이관된 과학데이터에 대한 체계화된 아카이빙 전략을 수립함으로써 과학데이터의 이관 및 관리의 용이성을 제공하여야 하며 특히, 과학데이터의 생산자의 소유권 등에 대한 고려를 충분히 인지할 수 있는 방안을 수립하여야 한다.

⑥번은 데이터 센터와 기관 리포지터리의 데이터전송에 있어서 국제 표준의 수용을 표현하고 있다. 생산된 과학데이터의 전송 및 수집정책에 있어서 국제표준의 준용은 체계화된 수집과 관련하여 매우 중요한 고려요소가 될 수 있다. 현재 정보자원의 전송 및 수집과 관련된 연구와 국제표준은 매우 활성화 되고 있다. 기존의 정형화된 정보자원과 과학데이터는 데이터의 형식적인 측면에서 유사하다. 따라서 정보유통과 관련된 검증된 국제표준의 적용을 통하여 과학데이터의 수집 및 유통이 활성화 될 수 있을 것으로 기대할 수 있다. 특히, 생산된 과학데이터와 메타데이터에 대한 국제표준의 적용은 향후 과학데이터의 재활용을 위한 접근성 제공에 있어서 매우 중요한 요소가 될 수 있다.

⑦번의 경우 출판사나 상용검색엔진에서 책의 출판과 웹상에서의 정보제공에 있어서 과학데이터와이 연계 또는 논문에서 과학데이터가 연계되어 서비스되는 모델을 나타낸다. 본 연구에서 제안하고 있는 과학데이터의 보존 및 활용을 위한 모델은 과학데이터의 재활용을 목표로 하고 있다. 따라서 외부 정보유통기관과의 협력은 과학데이터 활용의 활성화를 위하여 중요한 요소라고 할 수 있다. 특히, 웹 기반의 정보유통채널은 과학데이터의 유통에 직접적으로 적용이 가능하다. 따라서 기존의 정보유통기관으로서 다양한 포털 사이트와 대규모의 출판

사와의 협력은 필수적이라고 할 수 있다. 특히, Elsevier, EBSCO, Emerald 등의 대규모의 출판사들은 연구자들의 연구결과물을 출판에 있어서 과학데이터의 제출을 요구하는 추세라고 할 수 있다. 따라서 대형 출판사들은 과학데이터의 리포지터리로써 중요한 기관의 역할을 수행이 예상된다.

본 연구에서 제안하고 있는 모델은 과학데이터와 관련된 모든 요소와의 협력과 함께, 과학데이터의 관리체계 및 도구들을 고려하고 있다. 특히, 제안하고 있는 모델은 과학데이터와 관련된 국제적인 흐름을 최대한 수용하고 있으며 국내의 과학데이터와 관련된 환경적 요인들을 고려하였다.

6. 결 론

본 연구에서는 과학데이터와 관련하여 미국과 영국, 호주, 유럽연합의 국가별 동향을 조사하였으며, 해외 선진 사례의 경우 프로그램 단위별로 DataCite, ICSU의 WDS, PANGAEA, 하버드 대학의 Dataverse, BSRN, DLESE, IODP의 SEDIS, NASA의 GCMD에 대하여 조사 및 분석을 수행하였다. 해당 프로그램에 대한 조사/분석을 통하여 각각의 프로그램에서 공통된 시사점을 도출하였다. 이와 같은 분석결과를 기반으로 본 연구에서는 국내의 과학데이터 보존 및 활용을 위한 모델을 제시하였다. 과학데이터 관련 국내 연구 또는 서비스사례에 대한 조사를 수행하고자 하였으나, 국내에서는 관련된 사례가 미비하여 진행하지 못한 것은 아쉬운 점이다. 특히, 국내의 경우에 있어서 국가차원에서 지원되고 있는 다양한 연구들이 있음에도 불구하고 연구비 지원기관을 중심으로 과학데이터의 수집 및 관리 등과 관련된 정책이 전무하였다. 또한 국내에서 생산되는 과학데이터에 대한 체계적인 수집 및 관리를 통하여 통합적으로 서비스 제공 모델이 구현된 사례도 없다. 과학데이터는 국가 차원에 있어서 연구자들의 연구결과에 대한 검증과 함께, 과학데이터의 공유 및 재활용이라는 관면에서 경제적인 가치가 매우 크다고 할 수 있으며 이는 국가의 연구경쟁력 확보와 밀접하게 관련된다. 따라서 국가적인 과학데이터 관리정책 및 서비스 모델의 개발은 매우 중요한 현실적인 요구가 아닐 수 없다. 이와 같은 과학데이터와 관련된 제반 문제점과 관련하여 본 연구에서 조사 및 분석된 선진사례 및 프로그램과 함께, 제안하고 있는 과학데이터 관리 및 보존 모델은 기록물로서 과학데이터 보존과 활용을 위한 국가차원의 정책 수립에 있어서 기초자료로 활용될 수 있는 가치가 매우 높다고 할 수 있다.

본 연구에 대한 후속연구로서 과학데이터 관련 국가별 정책을 집중적으로 조사 분석하여 국내환경에 적합한 과학데이터 정책 수립에 지침이 될 연구는 체계화된 과학데이터관련 정책 수립에 필수적이라고 할 수 있다. 또한 과학데이터의 관리 및 활용을 위해서는 이질적 과학데이터에 대한 상호운용성(interoperability)을 확보할 수 있는 통합 메타데이터 스키마의 개발에 관한 연구가 요구된다. 특히, 유형, 특성, 형식 등에 있어서 서로 이질적인 과학데이터의 통합관리를 위한 메타데이터 체계의 개발은 과학데이터의 수집, 보존, 활용 등에 있어서 필수적인 도구로서 중요도가 매우 높다고 할 수 있다.

【참 고 문 헌】

박동진. 2010. 오스트레일리아의 과학데이터 서비스체제(ANDS) 구축과 시사점. 『KESLI 전자정보포럼 발표자료』.

박현영, 남태우. 2004. 디지털 아카이빙 정책에 관한 연구. 『제11회 한국정보관리학회 학술대회 논문집』, 69-76.

이수상. 2004. 디지털 아카이빙의 워크플로우와 보존처리 기술에 관한 연구. 『한국도서관·정보학회지』, 35(3): 119-138

정영임, 최호남, 최선희. 2010. 아카이빙 데이터의 활용성 증진을 위한 전략연구. 『한국정보관리학회지』, 27(1): 185-206.

정혜경. 2004. 디지털 아카이빙의 경제성 분석 연구. 『한국문헌정보학회지』, 38(4): 251-270.

채균식, 이응봉. 2003. 연구중에 생산된 과학기술 참조데이터 관리에 관한 연구. 『한국문헌정보학회지』, 37(4): 131-149.

CUL Data Working Group. 2008. Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library, US: Cornell University Library.

Shen, Danna. 2007. "Public sharing and understanding of scientific data - with the illustration of weather

terms." Data Science Journal(2007, 5), Volume 6, Supplement, 13.

Cheng, Jinpei. 2006. *Strategies for Preservation of and Open Access to Scientific Data in China: Summary of a Workshop.*

Koski, Kimmo, Claudio Gheller, Stefan Heinzel, Alison Kennedy, Achim Streit, and Peter WittenBurg. 2009. Strategy for a European Data Infrastructure. EU: PARADE.

Ruusalepp, Raivo. 2008. A comparative study of international approaches to enabling the sharing of research data, British: JISC.

The Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. 2009. Harnessing the power of digital data for science and society.

Hey, Tony, Steward Tansley, and Kristin Tolle. 2009. The Fourth Paradigm. Data-Intensive Scientific Discovery, Washington: Microsoft Research.

[인용 웹사이트]

Australian National Data Service, ANDS. [cited 2010.7.27]. 〈http://ands.org.au/about-ands.html〉.

DataCite. [cited 2010.6.21]. 〈http://www.datacite.org/〉.

Dataverse: An Open-Source Application for Publishing, Citing and Discovering Research Data. [cited 2010.5.25]. 〈http://thedata.org/〉.

Digital Curation Centre, DCC. [cited 2010.7.27]. 〈http://www.dcc.ac.uk/〉.

Digital Repository Infrastructure Vision for European Research, DRIVER. [cited 2010.6.20]. 〈http://www.driver-repository.eu/〉.

DLESE: Digital Library for Earth System Education. [cited 2010.4.10]. 〈http://www.dlese.org/〉.

GCMD: Global Change Master Directory. [cited 2010.3.7]. 〈http://gcmd.nasa.gov/〉.

GEWEX: Global Energy and Water Cycle Experiment. [cited 2010.5.12]. 〈http://www.gewex.org/〉.

ICSU. [cited 2010.6.21]. 〈http://www.icsu.org〉.

PANGAEA: Publishing Network for Geoscientific & Environmental Data. [cited 2010.5.25]. 〈http://www.pangaea.de/〉.

SEDIS: Scientific Earth Drilling Information Service. [cited 2010.3.5]. 〈http://sedis.iodp.org/〉.

WDC: World Data Center System USA Home. [cited 2010.6.21]. 〈http://www.ngdc.noaa.gov/wdc/〉.

WDS: World Data Systems. [cited 2010.6.19]. 〈http://www.icsu-wds.org/〉.

WRMC-BSRN: World Radiation Monitoring Center - Baseline Surface Radiation Network. [cited 2010.5.19]. 〈http://www.bsrn.awi.de/〉.