대학 내 연구자들의 연구데이터 관리에 관한 연구*

A Study on University Researchers' Data Management Practices

김 지 현(Jihyun Kim)**

〈 목 차 〉

초 록

본 연구는 대학 소속의 연구자들을 대상으로 연구데이터의 관리 현황을 분석하는 것을 목적으로 하였다. 이를 위해 2010년과 2011년 한국연구재단 지원사업의 최종선정과제 연구책임자들 중 대학에 소속되어 있는 이들을 대상으로 설문조사를 실시하였다. 총 131명의 설문 내용을 분석한 결과, 대학에서 생산·수집되는 연구데이터의 유형은 인문·사회과학 분야와 자연과학·의학 및 공학 분야 연구자들 간에 많은 차이를 보였다. 응답자들은 주로 관련정보와 데이터를 연결시키거나 추가적으로 생산된 데이터를 병합하는 활동을 통해 연구데이터의 활용 가치를 높이고 있었다. 연구데이터 보관을 위해 개인 PC나 이동식 매체가 보편적으로 사용되고 있었으며, 약 80%의 응답자들이 과제 종료 이후 연구데이터의 유용성이 유지되는 연한을 10년 미만으로 보았다. 연구데이터의 공유는 주로 소속 연구팀 내에서 이루어지거나 외부에서 데이터를 요청하는 연구자들과 이루어지고 있었다. 타인의 연구데이터를 활용하는 응답자의 비율은 본인의 데이터를 공유한다는 응답자의 비율보다 높았으며 다수의 응답자들이 출판된 논문에서 데이터를 추출하거나 개인적으로 연락하는 방식을 통해 타인의 연구데이터를 획득하고 있었다. 대학 내에서 연구데이터 관리체계가 존재하지 않는 경우가 대부분이었으며 특히 데이터의 장기보존과 메타데이터 작성에 대한 만족도가 낮아 이를 지원할 수 있는 연구데이터 아카이빙 서비스의 개발이 요청된다.

키워드: 연구데이터, 대학 연구자, 데이터 관리, 데이터 공유, 데이터 재이용

ABSTRACT

This study examined research data management practices from the perspectives of university researchers. A survey was conducted for principal investigators of projects in universities selected to be funded by National Research Foundation in 2010 and 2011. Predicated on the analysis of 131 survey responses, there was a great difference in types of research data between Humanities & Social Science fields and Science, Medicine and Engineering fields. Most respondents added value of their data by linking to other types of information or combining data from other sources. For storing data, PC or portable media were generally employed, and around 80% of respondents saw their data having a useful life under 10 years. Data was shared within research team and with outside researchers who requested data. The percentage of respondents who have reused data was higher than that of respondents who have shared data. In order to obtain data for reuse, the majority of respondents drew data from published articles, or contacted data creators. In most cases, mechanisms for managing data did not exist in projects or universities where respondents belong. Since the level of satisfaction with long-term preservation and metadata description of research data was found to be low, it was necessary to develop data archiving services to support the data management procedures.

Keywords: Research data, University researchers, Data management, Data sharing, Data reuse

* 이 연구는 2012년도 한국연구재단 신진연구자지원사업의 지원을 받아 수행된 연구임.

** 이화여자대학교 문헌정보학과 조교수(kim.jh@ewha.ac.kr)

• 접수일: 2012년 8월 28일 • 최초심사일: 2012년 9월 13일 • 최종심사일: 2012년 9월 21일

Ⅰ. 서 론

사이버인프라(cyberinfrastructure)와 e-Science, e-Research 등의 용어로 대표되는 21세기 첨단 과학기술 연구개발 환경은 데이터 집약적이고 정보 집약적인 분산적 협력 연구(distributed collaboration)에 중점을 두고 있다. 미국과학재단 (National Science Foundation)의 보고서에서는 사이버인프라의 핵심 요소로서 데이터를 관리, 보존하고 이를 다른 연구자들이 이용할 수 있도록 데이터 레포지터리를 구축하는 것을 제안하고 있으며 이를 지원하기 위한 비용과 전문화된 교육이 필요함을 강조하고 있다[1]. 영국의 e-Science 프로그램의 경우 자연과학, 공학, 의학 분야 100여개 연구 프로젝트로 구성되어 있으며 세부 분야별로 공통의 데이터 표준을 개발하거나 첨단 시뮬레이션 및 원격 시각화를 가능하게 하는 분산 어플리케이션을 개발하는 등 e-Science의 비전을 실현하기 위한 다양한 연구들이 진행되고 있다.[2]

데이터 관리와 보존의 중요성은 자연과학이나 공학에서뿐만 아니라 사회과학이나 인문학 분야에서도 강조되고 있다. 미국학술단체협의회 (American Council of Learned Societies)의 보고서에서는 사회과학 및 인문학 분야의 연구자들이 데이터를 생산, 보존하는 시스템에 의존하는 경향이 높아지고 있음을 언급하였다. 이를 지원하기 위하여 대학이나 대학 컨소시움에서 사회과학이나 인문학 관련 컴퓨팅 센터를 설립하고 데이터 관리에 대한 교육과 보존을 수행할 것을 제시하고 있다.[3] 영국의 경우 e-Social Science 센터가 설립되어 영국 경제사회연구협의회(Economic and Social Research Council)의 지원을 받아 운영되고 있다. 그리드를 기반으로 한 인프라를 구축하여 이를 통해 사회과학자들 간 연구수행에 필요한 데이터를 공유할 수 있게 한다는 목표를 지향하고 있다. 또한 디지털 인문학(digital humanities) 분야에서도 데이터 마이닝 기법을 활용한 연구나 문화 분석(cultural analytics)의 측면에서 방대한 양의 정량적 데이터를 이용하는 사례가 증가하고 있다.[4]

디지털 형태의 연구 데이터에 대한 관리와 보존 및 공유가 강조되고 있는 이유로 우선 데이터 공유를 통해 학자들이 좀 더 신속하게 연구문제를 탐구하고 해결할 수 있으며 이를 통해 학문 발전을 촉진시킬 수 있다는 점을 들 수 있다. 또한 데이터를 공개함으로써 중복연구를 방지할 수 있다는

---

1) National Science Foundation (NSF), *Revolutionizing Science and Engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.*(Arlington, VA: National Science Foundation, 2003), p.13.

2) Tony Hey, and Anne E. Trefethen, "Cyberinfrastructure and e-Science." *Science*, Vol.308(May 2005), p.819.

3) American Council on Learned Societies (ACLS). *Our Cultural Commonwealth: the Final Report of the American Council of Learned Societies commission on Cyberinfrastructure for the Humanities and Social Sciences.*(New Yok, NY: ACLS, 2006), pp.35.

4) Christine, L. Borgman, "The Digital Future is Now: A Call to Action for the Humanities," *Digital Humanities Quarterly*, Vol.3, No.4(2009). 〈http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html〉 [cited 2012.8.18]

장점이 있으며 연구자들은 연구비 수혜기관의 데이터 관리 정책에 따라 연구를 수행하게 된다. 뿐만 아니라 데이터 관리와 공유를 통해 연구 성과를 널리 알릴 수 있고 다른 연구자들에 의해 연구 결과의 재현과 확인이 가능하며 협동 연구를 증진시킬 수 있다. 이와 같은 기대효과에 힘입어 국가적 차원의 지원이 이루어지는 미국과 영국 등 제반 선진국에서 데이터 관리, 보존 및 공유에 대한 많은 연구들이 진행되어왔다.

국내에서는 한국과학기술정보연구원(KISTI)을 중심으로 2011년 11월 '국가과학데이터 심포지엄 2011'이 개최되면서 과학데이터의 공유에 대한 논의가 본격화되고 있다. 미국과 영국 등지에서 체계적인 과학데이터 관리와 공유가 진행되고 있는 반면 국내에서는 이에 대한 인식도가 매우 낮을 뿐만 아니라 과학자들이 원하는 인센티브가 존재하지 않는 한 과학데이터의 활발한 공유가 어려울 것이라는 전망이 제기되고 있는 실정이다. 이러한 논의 가운데서도 KISTI에서는 국가 과학데이터 공유·활성센터의 건립을 제안하고 데이터를 공유함으로써 얻을 수 있는 국가차원의 기대효과 - 새로운 융합연구 영역 개척, 시너지 창출, 산업발전에의 기여 - 를 제시하고 있다.[5)]

뿐만 아니라 국내 인문·사회과학 분야 연구데이터 아카이브로서 한국연구재단의 지원을 받아 운영되는 기초학문자료센터(Korean Research Memory: KRM)가 있다. KRM은 한국학술진흥재단과 연구재단의 지원을 받아 수행된 인문·사회과학 분야 연구들의 원 자료, 중간산출물 및 연구 성과물을 수집·제공하는 시스템이다. 국가가 지원하는 연구 사업의 성과물을 체계적으로 수집, 공유하고 학술적 가치가 높은 원 자료를 제공하여 후속 연구를 촉진시키며 국가 차원의 디지털 아카이빙 시스템을 구축한다는 목표로 2005년 설립되었다. 그러나 연구에 활용할 수 있는 데이터가 부족하고 문헌과 원 자료가 구분 없이 수록되어 있는 등 여러 가지 문제점들을 내포하고 있으며 이에 대한 개선책이 요구되는 실정이다.[6)]

본 연구에서는 국내외적으로 연구데이터의 체계적인 관리에 대한 요구가 높아지는 시점에서 연구데이터를 생산하는 대학 내 연구자들을 대상으로 데이터의 관리 및 공유 현황을 조사하는 것을 목적으로 하였다. 효과적인 연구데이터 아카이빙 서비스를 구축하기 위해서는 데이터를 생산하는 연구자들이 어떠한 방식으로 데이터를 관리하고 있고 필요사항이 무엇인지에 이해가 선행되어야 한다. 이를 위해 2010년과 2011년 한국연구재단 지원사업의 최종선정과제 연구책임자들 중 대학에 소속된 이들을 대상으로 설문조사를 실시하였다. 본 연구에서 살펴보고자 하는 연구 문제는 다음과 같다.

1) 대학 내 연구자들이 어떠한 종류의 연구데이터를 생산 혹은 수집하는가?
2) 대학 내 연구자들이 어떠한 방식으로 연구데이터를 관리하는가?

---

5) "과학데이터 공유가 새로운 경지를 연다" 대덕넷.
〈http://www.hellodd.com/kr/dd_news/article_view.asp?mark=36110〉 [cited 2012.8.18].
6) 신영란, 인문사회 연구데이터 아카이브의 발전방향에 관한 연구(석사학위논문, 이화여자대학교 정책과학대학원 기록관리 전공, 2012), pp.44-55.

3) 대학 내 연구자들의 연구데이터 공유의 범위와 정도는 어떠한가?
4) 대학 내 연구자들이 다른 연구자들의 데이터에 어떻게 접근하는가?

Ⅱ. 연구자들의 데이터 관리

1. 연구데이터 생애주기

연구데이터의 중요성이 부각되면서 데이터 생산에서부터 관리·보존·활용에 이르는 전반적인 과정을 생애주기(life cycle) 모형으로 개념화하고자 하는 논의가 진행되어왔다. 생애주기 모형은 데이터 혹은 정보가 생산되고 관리되는 단계들을 순차적으로 나타낸 것이며 연구수행과정의 일부로서 제안되고 있다.

연구 설계를 통해 데이터가 수집, 처리, 접근/배포, 분석되는 과정을 나타내고 있는 연구지식생산 생애주기 모형 (The Life Cycle of Research Knowledge Creation)에서는 연구자들이 문헌조사 등을 통해 연구주제를 개념화하는 단계에서부터 출발한다. 이후 연구비 신청을 위한 연구계획서를 작성, 제출하고 연구를 수행하며 이를 학회에서 발표하고 중간 보고서를 제출한 후 연구 성과를 학술지에 게재하고, 마지막에 이를 미디어를 통해 널리 알리거나 실제 업무에 활용하게 된다. 마지막 단계는 다시 처음 단계인 연구주제 개념화의 밑바탕이 되며 단계별로 순환하는 구조를 보여준다. 뿐만 아니라 데이터에 접근하거나 이를 배포할 때, 혹은 데이터 분석 시 기존의 데이터를 다른 목적으로 활용(Data Repurposing)하는 단계를 포함시켜 다양한 목적으로 연구데이터가 재이용(reuse)될 수 있음을 보여준다.[7)]

이처럼 연구와 학습의 과정은 본질적으로 순환하는 구조이고 그에 따른 연구 성과는 데이터와 정보를 바탕으로 생성되며 연구 성과를 통해 새로운 지식이 창출된다. 이를 반영한 학술지식주기(Scholarly Knowledge Cycle)에서는 효율적인 데이터 관리를 위한 요건으로 (1) 원본데이터의 무결성 유지; (2) 데이터의 출처(provenance) 개념에 대한 공통의 이해; (3) 표준에 근거한 원본데이터의 메타데이터 기술; (4) 원본데이터와 파생데이터 및 정보에 대한 식별이 제안되었다. 제시된 학술지식주기는 e-Bank UK 프로젝트의 개념적 모형으로서 연구데이터의 메타데이터 수집을 지원하는 시스템 구성 개발에 주안점을 두었다.[8)]

---

7) Charles Humphrey, e-Science and the life cycle of research, 2006. 〈http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc〉 [cited 2012.8.18].
8) Liz Lyon, "eBank UK: Building the Links between Research Data, Scholarly Communication and Learning," *Ariadne*, Vol.36, 2003. 〈http://www.ariadne.ac.uk/issue36/lyon/〉 [cited 2012.8.18]

초기의 연구들이 단선적인 생애주기 모형을 제안하였다면 최근 연구에서 제시되는 생애주기 모형은 다수의 하위주기를 포함하는 형태로 세분화된 양상을 보이고 있다. 영국 디지털 큐레이션 센터에서 제안하는 생애주기 모형은 다음의 4가지 측면으로 구성되어 있다: (1) 데이터 (데이터 객체 혹은 데이터베이스); (2) 전(全)생애 주기적 활동 (기술 정보; 보존 계획; 커뮤니티 감독 및 참여; 큐레이션과 보존); (3) 순차적 활동 (개념화; 생산·수령; 평가·선별; 입수; 보존활동; 저장; 접근·이용·재이용; 변용; (4) 비정기적 활동 (폐기; 재평가; 마이그레이션). 이러한 생애주기 모형은 보다 전문적이고 적극적인 연구데이터 서비스를 제공하기 위한 개념적 틀로서 활용될 수 있다.[9)]

2. 데이터 관리 및 공유에 대한 인식

데이터 생애주기 개념은 연구자들의 데이터 관리 현황을 조사한 기존 연구에서 활용되고 있는데, Carlson과 Anderson은 데이터 수집, 처리, 주석달기(annotation), 배포 및 재사용이라는 생애주기 4 단계를 중심으로 연구자들의 데이터 관리과정을 분석하였다. 이들은 자연과학 분야 2개 프로젝트와 설문데이터 분석을 위주로 하는 사회과학 분야 1개 프로젝트, 그리고 인류학 분야 1개 프로젝트를 대상으로 면담과 관찰을 통한 질적 연구를 수행하였다. 연구데이터 수집 시 인류학 분야의 경우 디지털 형태가 아닌 데이터에 대한 디지털화를 수행하는데 많은 비용이 소요되는 것으로 나타났다. 태생적 디지털 데이터를 생산하는 분야와는 달리 비(非)디지털 데이터가 상당량을 차지하는 분야일 경우 데이터베이스 작업이나 메타데이터 기술 등의 작업에 대한 비용이 높으며 이는 데이터 공유를 저해하는 요소로 작용할 수 있다. 데이터 처리의 경우 특히 정량적 연구 분야에서 데이터 시각화와 시뮬레이션을 위한 도구와 과정이 잘 정립되어 있으므로 데이터 재사용이 상대적으로 원활한 것으로 나타났다. 데이터 배포의 경우 데이터에 대한 소유권과 배포에 대한 동의 및 도덕적 권리에 대한 존중이 중요한 논의점으로 나타났다. 데이터 재사용에 있어서 저자들은 데이터의 수집 및 구조화에 대한 신뢰와 이를 다른 사람들에게 이해시키는 문제가 핵심임을 언급하였다. 또한 각 학문 분야의 역사와 연구자들이 속한 연구 커뮤니티의 특성이 데이터에 대한 맥락화와 기록화에 가장 큰 영향을 미치는 것으로 나타났다.[10)]

Tenopir와 동료 연구자들도 영국의 JISC(Joint Information Systems Committee) 에서 제안하는 데이터 생애주기 모형을 언급하면서 데이터 생산, 관리, 분석 및 공유는 연구수행에서 필수불가결한

9) Sarah Higgins, "The DCC curation lifecycle model," *International Journal of Digital Curation*, Vol.3, No.1(June 2008), p.136.

10) Samuelle Carlson, and Ben Anderson, "What are Data? The Many Kinds of Data and Their Implications for Data Re-Use," *Journal of Computer-Mediated Communication*, Vol.12, No.2(2007). 〈http://jcmc.indiana.edu/vol12/issue2/carlson.html〉 [cited 2012.8.18].

과정임을 설명하였다. 이들은 자연과학·의학 및 공학 분야 과학자들을 대상으로 국제적인 설문조사를 실시하였고, 1,329명의 설문 응답 분석을 바탕으로 데이터 공유의 장애요인과 공유를 촉진시킬 수 있는 조건을 논의하였다. 응답자들은 데이터 공유에 투자할 시간과 비용이 부족하다는 것을 가장 큰 장애요인으로 인식하고 있었으며 데이터의 장기적인 보존에 대한 현황에 만족하지 못하고 있는 것으로 나타났다. 대다수의 응답자들은 데이터에 대한 인용이나 협동연구의 기회를 얻는 것이 가능하다면 데이터를 공유하겠다는 의견을 보였다. 주된 연구비 지급기관, 학문 분야, 연령, 연구 초점 및 세계 권역(world regions)에 따라 데이터 관리 현황과 방식에 유의미한 차이가 있는 것으로 나타났다.[11)]

연구자들의 데이터 관리 현황을 분석한 또 다른 연구로서, 영국의 대학교육 기관들을 위한 국가 차원의 연구데이터 서비스인 UK Research Data Service(UKRDS)의 실현가능성을 평가한 연구가 있다. 이 연구에서는 다양한 방법론을 활용한 데이터 수집이 이루어졌는데 브리스톨(Bristol), 리즈(Leeds), 레스터(Leicester), 옥스퍼드(Oxford) 대학들과 협력하여 이들 대학의 연구자들을 대상으로 설문조사, 면담, 포커스 그룹 및 워크샵을 수행하였다. 연구자들이 속한 학문분야에 따라 데이터 관리에 대한 의무사항과 방식에 차이를 보였으며 학과나 개인 차원에서 데이터를 관리하는 것이 일반적인 것으로 나타났다. 이러한 로컬 차원에서의 관리를 넘어 장기적인 데이터 보존을 위한 데이터 레포지터리의 필요성이 제안되었으며, 데이터 관리에 대한 기관의 지원과 이용 가능한 데이터 서비스에 대한 정보 제공 및 데이터 보안과 접근 통제에 대한 요구도 함께 제시되었다.[12)]

영국의 Research Information Network(RIN)에서도 8개 학문분야 연구자들을 대상으로 데이터의 생산 및 관리 과정을 조사하였는데 연구자들의 태도 및 요구사항과 데이터 관리를 위한 인프라 및 정책에서 학문분야별로 많은 차이를 보였다. 잠재적인 가치를 지닌 데이터 다수가 효율적으로 관리되지 못하고 있었으며, 연구비 제공기관의 데이터 관리 정책이 학문 분야의 통념이나 실제와 맞지 않는 경우도 있었다. 메타데이터의 제공 양상과 수준은 표준 메타데이터에서부터 연구자들이 임의로 제공하는 레이블에 이르기까지 다양한 것으로 나타났다. 또한, 데이터에 부가가치를 제공하는 방법으로 주석을 제공하고, 추가적인 데이터를 기존 데이터에 병합하거나 데이터 접근 및 분석을 위한 도구를 개발하는 등 여러 가지 방식들이 활용되고 있었다.[13)] 연구자들은 데이터 공유를 통해 학문의 발전을 도모할 수 있고 긍정적인 피드백이나 존경, 혹은 공저자나 협력연구의 기회를 확보할 수 있다는 점을 장점으로 인식하고 있었다. 그러나 시간 부족, 데이터 관리에 대한 경험이나 지식

11) Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions," *PLoS ONE*, Vol.6, No.6(June 2011). 〈http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101〉 [cited 2012.8.18].

12) Neil Beagrie, Robert Beagrie, and Ian Rowlands, "Research Data Preservation and Access: The Views of Researchers," *Ariadne*, Vol.60(2009) 〈http://www.ariadne.ac.uk/issue60/beagrie-et-al/〉 [cited 2012.8.18]

13) Alma Swan, and Sheridan Brown, *To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network*, 2008. 〈http://eprints.soton.ac.uk/266742/〉 [cited 2012.8.18].

부족, 법적·윤리적인 문제, 연구 아이디어 표절에 대한 우려 등으로 인해 데이터 공유를 꺼리는 경향을 보였다.[14]

이렇게 다양한 분야의 연구자들을 대상으로 데이터 관리와 공유에 대한 인식을 조사한 연구뿐만 아니라 특정 분야 과학자들을 대상으로 한 연구들도 다수 존재한다. 생태학 분야의 학제적 협력연구 네트워크인 Long-Term Ecological Research(LTER)에 속한 연구자들의 데이터 관리와 공유 행태를 조사한 연구에서 생태학자들은 데이터 공유의 필요성을 인식하면서도 학술지 논문을 출판하기 이전의 데이터 공유는 매우 부담스럽고 시간이 많이 소비되며 그에 대한 보상이 없다는 것을 문제점으로 지적하였다.[15] 이와 유사하게 미국과학재단 과학기술센터 중 하나인 CENS(The Center for Embedded Networked Sensing)의 과학자들 역시 출판 예정인 데이터보다 공식적으로 출판된 데이터를 공유하려는 성향이 더 강한 것으로 나타났다.[16]

데이터의 재이용에 대한 연구자들의 인식을 조사한 Zimmerman은 생태학자들을 대상으로 자신들이 수집하지 않은 데이터를 재이용할 때 어떻게 그 품질을 평가하는지를 분석하였다. 그는 데이터 수집방법의 표준화가 데이터 재사용을 촉진하는데 도움이 되며 이를 통해 로컬 단계의 과학지식이 공공의 영역으로 확장될 수 있음을 주장하였다. 뿐만 아니라 로컬 단계에서의 연구 맥락을 이해하는 것이 데이터 재사용에 있어 매우 중요한 요소임을 강조하였다.[17] 이러한 연구맥락의 이해와 관련하여 지진공학 연구자들을 대상으로 한 조사에서는 데이터를 재이용하기 전에 연구자들이 그 데이터의 적합성을 평가하고 데이터가 어떠한 실험을 통해 생산되었는지를 자세히 이해하는 것이 중요하다고 하였다. 또한 연구자들이 반드시 해당 데이터를 신뢰할 수 있어야 하며 이를 위해서는 데이터에 대한 적절한 메타데이터 제공이 필요하다는 점이 강조되었다.[18]

데이터의 장기보존을 위해서도 데이터 생산자가 아카이브에 데이터를 제공하기에 앞서 데이터에 대한 충분한 메타데이터를 제공하는 것이 중요하다. 그러나 실제 연구자들의 데이터 관리에서 데이터에 대한 맥락 정보를 제공하고 메타데이터를 기술하는 작업이 우선순위가 높은 다른 업무에 밀리는 경우가 많다.[19] 이러한 문제에 초점을 맞추어 Hedstrom과 Niu는 미국 법무부 국립연구소

14) Aaron Griffith, "The Publication of Research Data: Researcher Attitudes and Behaviour," *International Journal of Data Curation*, Vol.1, No.4(2009), p.51-52.

15) Helena Karasti, Karen S. Baker, and Eija Halkola, "Enriching the Notion of Data Curation In E-Science: Data Managing and Information Infrastructuring in the Long-Term Ecological Research (LTER) Network," *Computer Supported Cooperative Work*, Vol.15(2006), p.327.

16) Christine Borgman, Jillian C. Wallis, and Noel Enyedy, "Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology," *Lecture Notes in Computer Science*, No.4172(2006). 〈http://www.springerlink.com/content/6lx0112538724ql4/〉 [cited 2012.8.18].

17) Ann S. Zimmerman, "New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data," *Science, Technology & Human Values*, Vol.33, No.5(2008), p.649.

18) Ixchel M. Faniel & Trond E. Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data," *Computer Supported Cooperative Work*, Vol.19(2010), p.369-372.

19) Angus Whyte, Dominic Job, Stephen Giles, and Stephen Lawrie. Meeting Curation Challenges in a Neuroimaging

(National Institute of Justice)에서 연구비 지원을 받은 55명의 사회과학 분야 연구자들을 대상으로 데이터의 메타데이터 기술에 대한 설문조사를 실시하였다. 설문응답자들은 자신들이 수집한 데이터가 어떻게 활용되고 이를 통해 일반대중들에게 어떤 혜택이 제공되는지를 알게 된다면 보다 기꺼이 데이터에 대한 메타데이터를 기술하고 이를 데이터 아카이브에 제공하겠다고 응답하였다. 그러나 이들은 연구 결과물을 출판하기 전에 연구데이터를 공개하는 것과 데이터에 대한 비밀유지의 어려움 및 데이터에 대한 통제권 상실을 우려하기도 하였다. 이러한 비밀유지와 윤리적인 측면에 대한 우려는 캐나다 사회과학 인문학 연구 협의회 (Social Science and Humanities Research Council: SSHRC)로부터 연구지원비를 받은 연구자들을 대상으로 조사한 연구에서도 드러난다.[20] Hedstrom과 Niu는 연구데이터의 원활한 수집과 보존을 위해서는 데이터 아키비스트와 연구자들 간의 협력이 매우 중요하며 연구자들에 대한 인센티브를 개발하는 것이 필요함을 강조하였다.[21]

국내의 연구자들을 대상으로 연구데이터의 제출과 공유에 영향을 미치는 요인을 조사한 연구에서는 정보인프라, 동기부여, 장애요인의 세 가지로 요인 분석이 이루어졌다. 정보인프라 요인에는 연구데이터 관리시스템, 메타데이터, 연구데이터 표준화, 데이터 관리지원 및 교육 프로그램이 포함되었다. 동기부여 요인은 내적요인과 외적 요인으로 구분되며 내적 요인으로 데이터의 공유가 다른 연구자들에게 도움이 된다는 인식, 과제완료에 대한 안정감, 소속기관 직원으로서의 의무감, 만족감이 제시되었다. 외적요인은 데이터제출 의무화, 연구실적 인정, 금전적 보상 및 좋은 평판으로 요약되었다. 장애요인으로는 추가적인 시간과 노력의 문제와 데이터에서 더 많은 논문을 출판하려는 연구자들의 바램, 데이터를 공유함으로써 이를 독점적으로 사용할 수 없다는 점과 기밀성을 유지하기 어렵다는 점이 포함되었다. 2010년 SCI 논문을 출판한 국내 정부출연연구기관 소속 연구자들을 대상으로 설문조사를 통해 분석한 결과 내적동기부여요인이 데이터 제출의도에 가장 큰 영향을 미치는 것으로 나타났다. 이를 통해 내적동기부여요인을 강화하는 것이 연구데이터 제출을 독려하는데 가장 효과적인 방법임을 제안하였다.[22]

Group. *International Journal of Digital Curation*, Vol.3, No.1(2008), p.174.

20) Carol Marie Perry, "Archiving of Publicly Funded Research Data: A Survey of Canadian Researchers," *Government Information Quarterly*, Vol.25(2008), p.145.

21) Margaret Hedstrom, and Jinfang Niu, "Incentives for Data Producers to Create "Archive-Ready" Data: Implications for Archives and Records Management," *Proceedings of 2008 Society of American Archivist Research Forum*. 〈http://www2.archivists.org/sites/all/files/M-HedstromJ-Niu-SAA-ResearchPaper-2008.pdf〉[cited 2012.8.18.]

22) 김은정, 연구데이터 수집에 영향을 미치는 요인 분석(박사학위논문. 중앙대학교 대학원 기록관리학과 기록물관리학 전공, 2012), p.94-96.

Ⅲ. 연구방법 및 설문응답자 분석

본 연구의 데이터 수집을 위해 설문조사법이 활용되었으며 온라인 설문도구인 서베이몽키(www.surveymonkey.co.kr)를 사용하여 설문을 실시하였다. 설문 대상은 2010년과 2011년 한국연구재단 지원사업의 최종선정과제 연구책임자들 중 대학에 소속되어 있는 연구자들로 한정하였다. 한국연구재단은 교육과학기술부의 연구 사업을 통합·지원하는 대표적인 연구관리전문기관으로 다양한 학문분야 연구과제에 대한 연구지원을 수행하고 있다. 한국연구재단의 연구지원을 받는 대학 내 연구자들은 활발한 연구 및 저술 활동을 통해 연구데이터를 생산, 수집하고 있을 것이라 가정할 수 있으므로 이들을 조사 대상으로 삼았다.

설문 대상자 모집단을 파악하기 위해 한국연구재단 웹사이트의 '전체사업공지' 게시판에서 2010년과 2011년에 공지된 최종선정과제들의 리스트를 수집하였다. 먼저 게시판의 검색기능을 활용하여 '최종 선정' 혹은 '선정'이라는 키워드로 검색된 결과물 가운데 2010년과 2011년에 해당하는 것으로 범위를 좁혔다. 이렇게 검색된 리스트에는 연구책임자 이름, 소속기관, 학문분야, 과제명. 지원액, 연구기간이 명시되어 있었으며 이를 엑셀 파일에 수록하였다. 공지된 사업 중에서 시간강사 지원사업과 같이 연구수행활동과 직접적인 관련성이 낮은 사업이나 연구책임자 성명이 표시되지 않은 사업들은 제외하였다. 결과적으로 대학에 소속된 연구자들만을 포함하는 모집단은 인문·사회분야 4,601명과 자연과학·공학 분야 8,383명으로 모두 12,984명인 것으로 나타났다.

본 연구에서는 이들 모집단의 10%를 무작위 추출하였고 온라인 설문 배포를 위해 추출된 연구자들의 소속대학 홈페이지에 나와 있는 이메일을 수집하였다. 이 과정에서 이메일 주소가 확인된 연구자들이 인문·사회분야 348명과 자연과학·공학 분야 680명인 것으로 나타났다. 이들 중 2010년과 2011년에 중복되어 있는 연구자들과 서베이몽키 이메일을 수신 거부하는 연구자들을 제외하고 인문·사회분야 342명, 자연과학·공학 분야 675명인 것으로 나타나 설문 표본을 총 1,017명으로 확정지었다.

설문지는 (1) 데이터의 유형 2문항; (2) 데이터의 관리 및 공유 현황 9문항; (3) 데이터 관리과정의 만족도 6문항 (5점 척도); (4) 데이터 공유 및 공개에 대한 인식 28문항 (5점 척도); (5) 응답자 개인특성 6문항으로 구성되었다. 설문문항 작성을 위해 설문조사를 수행한 선행연구의 문항들을 수정하여 설문지에 포함시켰다.[23] 또한, 일부 문항들은 관련 연구의 내용을 기반으로 새롭게 작성하였다.[24] 서베이몽키를 통해 작성된 설문지는 이메일로 배포되었으며, 설문의 목적에 대한 설명과 설문지 링크, 응답에 대한 사례로 백화점 상품권 1만 원 권을 우송한다는 내용을 포함하였다. 본 연구에서는 설문지 항목 중 관리현황에 해당되는 내용인 데이터의 유형, 관리 및 공유 현황과 관리

23) 김은정, 전게서; Beagrie, Beagrie and Rowlands, *op.cit.*; Tenopir, Allard, Douglass et al., *op.cit.*

24) Christine, L. Borgman, "Data, Disciplines, and Scholarly Publishing," *Learned Publishing*, Vol.21(2008); Carlson and Anderson, *op.cit*; Swan and Brown, *op.cit.*

과정의 만족도에 대한 항목들에 대한 데이터를 분석하였다.

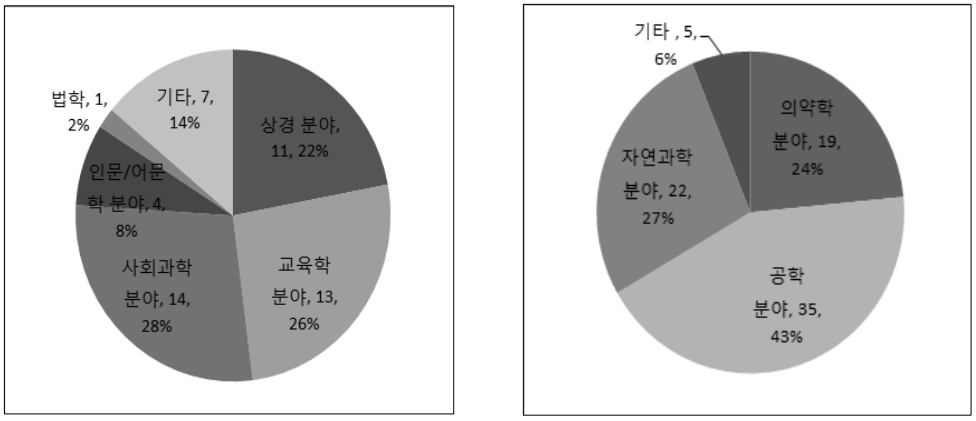
설문 데이터는 2012년 8월 2일부터 17일 사이에 수집되었으며, 설문지의 80% 이상이 응답된 131부를 분석 대상으로 삼았다. 응답률은 약 13%로 나타났으며 낮은 응답률로 인해 본 연구에서 수집된 데이터로 모집단의 대표성을 논의할 수 없다는 한계가 존재한다. 따라서 설문 분석 결과를 통해 나타난 대학 내 연구자들의 데이터 관리 행위의 일면을 분석하고 이에 대한 시사점을 도출하고자 한다.

우선 응답자들의 직위 분포를 살펴보면 정교수가 절반을 넘는 74명(56.5%)인 것으로 나타났으며 조교수인 응답자들이 그 다음으로 많은 수를 차지했다 (〈표 1〉 참조). 소수이기는 하지만 연구교수 또는 초빙교수인 응답자들과 책임연구원 또는 박사후연구원으로 대학에 소속되어 연구를 수행하는 응답자들도 있었다. 또한 응답자들 중 103명(78.6%)이 남성, 나머지 28명(21.4%)은 여성인 것으로 나타나 응답자들의 대다수가 남성임을 알 수 있다.

〈표 1〉 응답자의 직위

	빈도	퍼센트
정교수	74	56.5
부교수	14	10.7
조교수	36	27.5
연구교수	4	3.1
초빙교수	1	0.8
연구원	2	1.5
합계	131	100.0

응답자들의 40.5%에 해당하는 53명은 인문·사회분야 지원사업의 연구비 수혜자인 반면 나머지 78명(59.5%)은 자연과학·의학 및 공학 분야 지원사업의 연구비 수혜자인 것으로 나타났다. 그러나 인문·사회분야 지원사업 수혜자인 응답자들 중 5명은 의학, 물리학 또는 공학 분야의 연구자들인 것으로 나타났고 이공계 분야 지원사업 연구비 수혜자 중에서도 행정학 분야 연구자 2명이 포함되



〈그림 1〉 연구 분야 분포: 인문 · 사회과학 분야 응답자(좌) & 자연과학 · 의학 및 공학 분야(우)

어 있었다. 결과적으로 응답자들 중 인문·사회분야 연구자들은 50명, 자연과학·의학 및 공학 분야 연구자들은 81명인 것으로 나타났다. 각 집단의 연구 분야 분포는 〈그림 1〉에서 제시되어 있다.

인문 사회과학 분야 응답자들 중에서 경영학, 국제통상 및 무역에 해당하는 상경분야, 교육학, 아동학 및 영어교육, 국어교육 등과 같은 사범계열 분야를 포함하는 교육학 분야, 사회학, 심리학, 정치외교학, 사회복지학 등을 포함하는 사회과학 분야의 연구자들이 대다수를 차지하고 있음을 알 수 있다. 그러나 인문학이나 어문학 계열의 응답자는 소수인 것으로 나타났다. 이에 비해 이공계 응답자들 중에서는 전자공학, 기계공학, 컴퓨터공학 등 다양한 분야의 공학 연구자들이 43%를 차지하고 있었다. 물리학, 화학, 생명과학과 같은 자연과학 분야와 의학, 한의학, 약학 및 간호학을 포함하는 의약학 분야의 연구자들이 비슷한 숫자로 응답하였다. 기타에 해당하는 인문 사회과학 분야들은 관광경영, 호텔경영, 도시계획, 의상학 등이었으며, 이공계 분야로는 식품영양학, 물리치료, 과학교육 등이 포함되었다.

Ⅳ. 연구데이터 관리 현황

1. 데이터의 생산 및 수집데이터의 유형

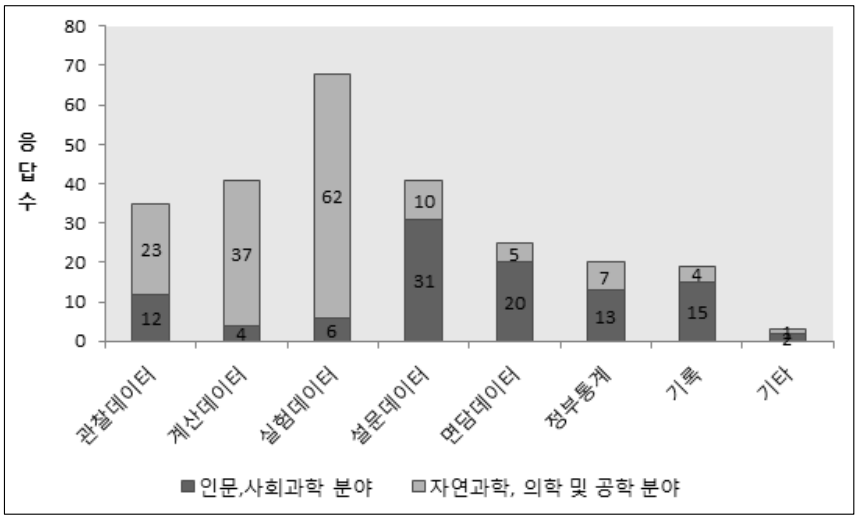
응답자들이 연구수행 과정에서 생산 또는 수집하는 데이터의 유형을 조사한 결과 실험실 환경, 실험장비 혹은 현장실험에서 생산되는 실험데이터에 대한 응답이 가장 많았다(〈표 2〉 참조). 컴퓨터 모델링이나 시뮬레이션을 통해 생산되는 계산데이터와 설문데이터에 대한 응답 수가 그 다음으로 많았으며, 특정한 장소나 시간에서 이루어지는 관찰을 통해 생성되는 관찰데이터를 생산하는 연구자들이 세 번째로 많은 것으로 나타났다. 면담데이터나 정부통계 뿐만 아니라 신문, 사진, 서신, 서적, 일기, 필름, 지도 등과 같은 기록을 연구의 원 자료로서 활용하는 연구자들도 상당수 존재했다. 기타 데이터의 유형으로 질병관리본부 주관의 국민건강영양조사 자료, 웹 데이터 (텍스트, 하이퍼링크, 이미지 등) 혹은 기업의 재무 데이터를 수집한다는 의견이 있었다.

〈표 2〉 연구데이터의 유형*

	빈도	퍼센트
실험데이터	68	27.0
계산데이터	41	16.3
설문데이터	41	16.3
관찰데이터	35	13.9
면담데이터	25	9.9
정부통계	20	7.9
기록	19	7.5
기타	3	1.2
합계	252	100

*복수응답 허용

Borgman이 언급한 바와 같이 학문 분야 별로 생산, 수집되는 데이터의 유형에는 차이가 있으며 자연과학 분야에서 활용되는 데이터는 연구를 목적으로 생산된 것이 대부분인데 반해 인문학이나 사회과학 분야 데이터는 본래 연구를 목적으로 생산되지 않은 데이터를 활용하는 경우가 많다는 점에서 차이가 있다고 하였다.[25] 〈그림 2〉는 응답자를 인문·사회과학 분야와 자연과학·의학 및 공학 분야의 두 집단으로 나누어 데이터의 유형을 살펴본 결과를 나타낸 것이다.



〈그림 2〉 분야 별 생산 · 수집되는 데이터 유형의 차이

그림에서 제시된 바와 같이 실험, 계산 및 관찰 데이터는 자연과학·의학 및 공학 분야에서 주로 생산된다는 것을 알 수 있다. 실험데이터와 계산데이터의 경우 90% 이상의 응답자들이, 그리고 관찰데이터는 66%의 응답자들이 이공 및 의학 분야 연구자들인 것으로 나타났다. 그러나 설문, 면담 데이터, 정부통계 및 기록의 경우 인문·사회과학 분야 연구자들이 주로 생산, 수집하는 데이터 유형인 것으로 드러났다. 설문데이터의 경우 계산데이터와 함께 두 번째로 높은 응답을 보인 데이터 유형이었는데 76%에 해당하는 31명이 인문·사회과학 분야 소속인 것으로 나타났다. 면담 데이터와 기록은 80% 가량의 응답자들이 인문·사회과학 분야인 것으로 나타났고 정부통계를 수집하는 연구자들도 이 분야의 연구자들이 대다수인 것으로 나타났다.

2. 데이터의 처리 및 보관

Swan과 Brown은 연구자들이 데이터를 수집한 이후 데이터의 검색이나 재이용을 용이하게 하거

25) Borgman, *op.cit.*, pp.31-32.

나 내용을 풍부하게 하고, 접근성을 향상시키는 모든 행위를 가치를 부여하는 활동(adding value)으로 정의하였다. 데이터를 정리, 확인, 조직화하면서 이에 대한 메타데이터를 부여하고 주석을 작성하며 관련 정보들과 연결시키는 기록화 작업들이 대규모 데이터 센터에서는 보다 체계적으로 이루어질 수 있다. 그러나 연구자 개인 차원에서 이루어지는 데이터 처리는 임시방편적으로 이루어지는 경우가 대부분이다.[26]

응답자들이 어떠한 방식으로 데이터의 활용 가치를 높이는 작업을 수행하고 있는지를 〈표 3〉을 통해 확인할 수 있다. 절반 이상을 차지하는 응답이 논문이나 기타 관련 정보들과 데이터를 연결하는 것과 연구과정 중에 생성된 추가적인 데이터를 병합하는 활동이었는데 이는 연구수행 과정에서 자연스럽게 진행될 수 있는 특징을 가진다. 반면에 데이터에 대한 주석을 작성한다든지 데이터 분석이나 조작을 위한 도구를 개발하는 등의 추가적인 데이터 처리 작업에 대한 응답은 상대적으로 적었으며 메타데이터를 작성한다는 응답이 가장 적은 것으로 나타났다. 기타 사항으로 데이터 수집 등에 대한 노트를 작성한다는 연구자 1인이 있었으므로 이를 응답에 포함시켰다.

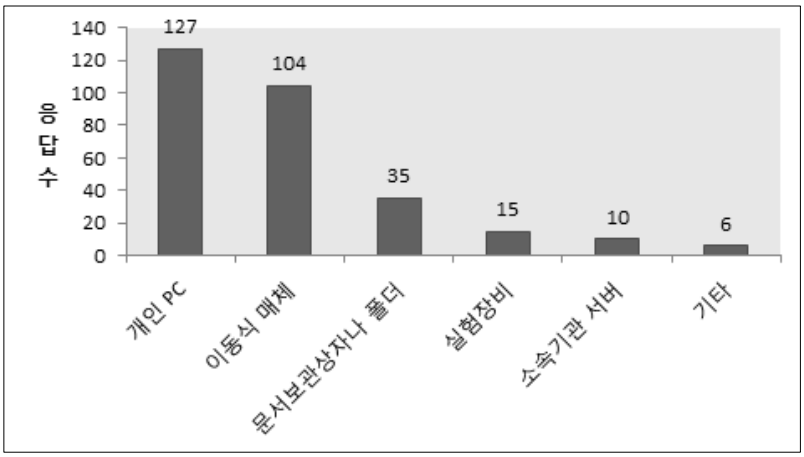
〈표 3〉 연구데이터의 처리 방식*

	빈도	퍼센트
논문이나 기타 관련 정보들과 데이터를 연결	96	34.4
관련 실험이나 연구에서 생산된 추가적인 데이터를 병합	80	28.7
데이터의 생산 환경이나 맥락 등에 대한 주석 작성	41	14.7
데이터를 조작/분석하는 도구 개발	34	12.2
데이터의 검색이나 체계화를 위한 관리항목(메타데이터) 작성	20	7.2
데이터 수집 등에 관한 노트 작성	1	0.4
해당사항 없음	7	2.5
합계	279	100.0

*복수응답 허용

데이터의 보관 장소와 관련하여 연구자들은 주로 개인 PC나 USB, CD, DVD와 같은 이동식 매체를 사용하는 것으로 나타났다(〈그림 3〉 참조). 복수응답을 허용하여 모두 297개의 응답 중에 127(42.8%)명이 개인 PC를, 104명(35.0%)이 이동식 매체를 이용하는 것으로 나타나 단기적인 데이터 활용을 위한 보관이 일반적인 것으로 나타났다. 물리적인 매체에 수록된 데이터를 보관하는 방식으로 문서보관상자나 폴더를 이용한다는 응답도 있었다. 기타 방식으로 4명의 응답자가 웹 하드 또는 클라우드 시스템을 언급하였으며, 구글 메일을 데이터 보관 장소로 활용한다는 연구자도 있었다. 이처럼 다양한 방식으로 디지털 형태의 연구데이터가 보관되고 있으나 사용되는 방식들이 장기적인 보존의 측면에서는 취약하다는 것을 알 수 있다.

26) Swan and Brown, *op.cit.*



〈그림 3〉 연구데이터 보관 장소

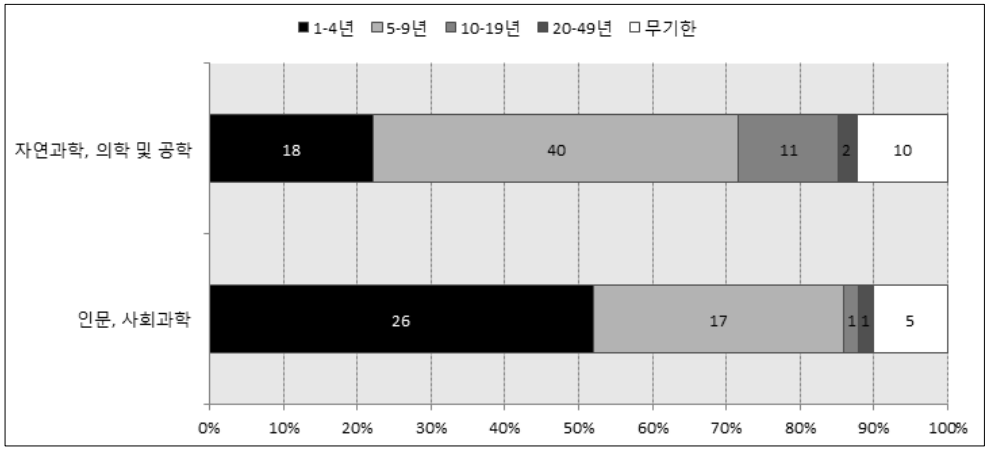
연구과제가 종료된 이후 어느 정도의 기간 동안 연구데이터가 유용할지에 대한 질문에서 5년에서 9년 사이라고 응답한 이들이 전체의 43.5%를 차지하였으며 그보다 짧은 1년에서 4년 사이라고 응답한 사람들이 33.6%인 것으로 나타났다(〈표 4〉 참조). 영국 UKRDS 설문조사 연구에서는 같은 질문에 대하여 48.9%의 응답자들이 10년 미만을 데이터의 유용성이 존재하는 기간이라고 응답[27]한 반면 본 연구에서는 78%가 응답해 큰 차이를 보였다. 10년 이상 장기보존을 원하는 응답자들도 소수 존재하였으며 특히 연구데이터의 유용성을 무기한으로 보는 응답자들도 있었다.

〈표 4〉 데이터가 유용성을 가지는 연한에 대한 인식

	빈도	퍼센트
1-4년	44	33.6
5-9년	57	43.5
10-19년	12	9.2
20-49년	3	2.3
무기한	15	11.5
	131	100.0

〈표 4〉에서 제시된 응답을 인문사회 분야와 이공 분야 응답자 두 집단으로 구분하여 보았을 때 인문사회과학 분야 응답자의 절반 이상인 26명의 응답자들이 1년에서 4년으로 응답한 반면 자연과학·의학 및 공학 분야 응답자들의 거의 절반인 40명의 응답자들이 5년에서 9년으로 응답하였다(〈그림 5〉 참조). 본 연구의 응답자 중 인문, 어문학 또는 예술 분야 연구자들이 거의 포함되어 있지

27) Beagrie, Beagrie and Rowlands, *op.cit.*



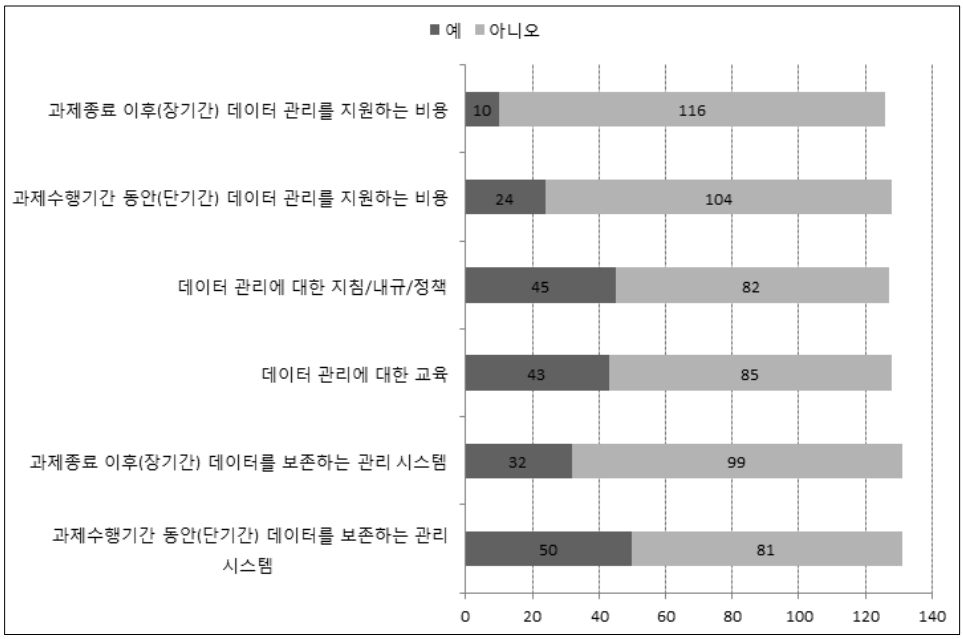
<그림 5> 분야 별 데이터의 유용성이 유지되는 연한에 대한 인식

않고 대다수 상경 및 사회과학 분야 연구자들임을 감안할 때 중단기적인 데이터 보존에 대한 요구가 10년 이상 장기적인 보존에 대한 요구보다 높은 것으로 보인다. 자연과학·의학 및 공학 분야 응답자들 중에는 10년 이상 장기적인 보존을 원하는 연구자들이 인문·사회과학 분야 연구자들보다 많다는 것을 알 수 있다. 영국 UKRDS 설문조사에서는 공학 및 자연과학 분야 또는 교육학 분야 연구자들이 대체적으로 10년 미만의 기한을 선택한 반면 경영학 분야는 10년-19년, 지리학, 임상의학. 보건학 및 예술 분야 연구자들은 10년 이상 또는 무기한을 선택한 경우가 많았다.[28]

연구 프로젝트나 학과 또는 실험실 등에서 연구데이터의 보존 및 관리에 대한 체계가 존재하는지에 대한 질문에서 과제수행기간 동안, 즉 단기간 데이터를 보존하는 관리 시스템이 존재한다는 응답자가 50명(38.2%)으로 가장 많았다(<그림 6> 참조).

데이터 관리에 관한 지침이나 내규 또는 정책이 존재한다는 응답과 데이터 관리에 대한 교육을 실시한다는 응답이 그 다음으로 많아 각각 35.4%와 33.6%를 차지하였다. 데이터 관리를 장/단기적으로 지원하는 비용이나 장기간 데이터를 보존하는 관리 시스템에 대한 응답 수는 상대적으로 적은 것으로 나타났다. <그림 6>에서 제시된 바와 같이 사실상 절반이 넘은 응답자들이 데이터 관리 체계를 갖추고 있지 않은 것으로 나타나 대학 내에서 체계적인 연구데이터 관리가 이루어지고 있지 않다는 것을 알 수 있다.

28) *Ibid.*



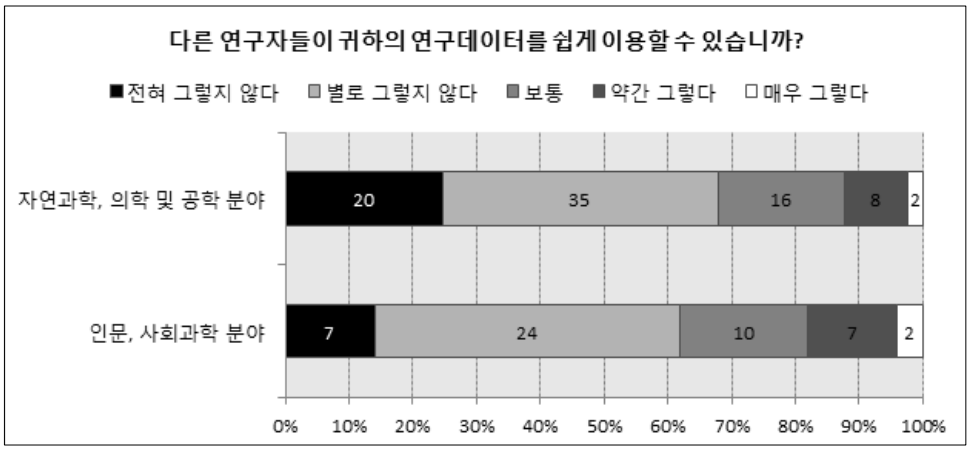
〈그림 6〉 데이터 관리체계 유무

3. 데이터 공유와 재이용

대학 내 연구자들의 데이터 공유 정도를 조사하기 위해 그들의 데이터를 다른 사람들이 쉽게 이용할 수 있는지를 5점 척도에 표시하게 하였다. 〈그림 7〉은 수집된 응답들을 인문·사회과학 분야와 이공·의학 분야의 두 집단으로 구분하여 나타낸 것이다. 두 집단 모두 60% 이상의 응답자들이 다른 연구자들이 응답자의 데이터를 쉽게 이용할 수 없다고 응답하였다. 쉽게 이용할 수 있다고 응답한 비율은 두 집단 모두 20% 미만인 것으로 나타나 데이터를 다른 연구자들이 이용하기 용이하게 공유 또는 공개하는 정도는 낮은 것으로 나타났다.

응답자들의 데이터 공유의 범위는 〈표 5〉에 제시되어 있는데, 데이터를 응답자 본인만 이용하고 타인과 전혀 공유하지 않는 응답자가 4명(3.1%)인 것으로 나타났다. 데이터를 타인과 공유하는 경우로 우선 소속 연구팀의 연구자와만 데이터를 공유한다는 응답자와 더불어 소속 연구팀의 연구자 또는 데이터를 요청하는 외부 연구자와 공유한다고 응답한 사람이 44명(33.6%)인 것으로 나타났다. 이들은 연구자들과만 데이터를 공유하는 경우라고 볼 수 있다.

이보다 공유 범위를 확장하여 소속 연구팀 내에서만 모든 사람들과 공유한다는 응답은 32명(24.4%)이었고 소속 연구팀원들 뿐만 아니라 외부 연구자들이 요청할 경우 공유한다는 응답자는 37명(28.2%)이었다. 데이터를 요청하는 외부 연구자들과만 공유한다는 응답도 소수 존재했다. 가장



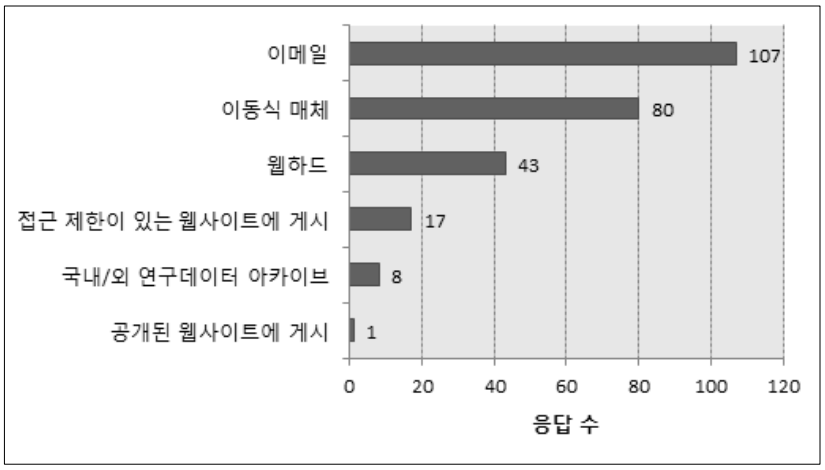
〈그림 7〉 데이터의 공유 정도

넓은 공유 범위로서 오픈 액세스로 모든 사람들이 활용할 수 있게 공개한다는 응답자는 7명(5.3%)인 것으로 나타났다. 기타 의견으로는 응답자가 초청강의를 할 때 주최 측에서 수업자료가 필요하다고 요구하는 경우 연구데이터 중에 이미 출판된 데이터는 공유한다는 응답이 있었다. 또한, 소속 연구원이나 외부에서 요청하는 연구자 중에서 허락된 사람들에 한해서 공유한다는 의견도 있었다.

〈표 5〉 데이터의 공유 범위

	빈도	퍼센트
응답자 본인만 이용	4	3.1
소속 연구팀의 책임 · 공동 연구자와만 공유	34	26.0
소속 연구팀의 책임 · 공동 연구자 & 데이터를 요청하는 외부 연구자와 공유	10	7.6
소속 연구팀의 모든 사람들과만 공유	32	24.4
소속 연구팀의 모든 사람들 & 데이터를 요청하는 외부 연구자와 공유	37	28.2
데이터를 요청하는 외부 연구자들과만 공유	5	3.8
모든 사람들 (데이터를 누구나 활용할 수 있게 공개)	7	5.3
기타	2	1.5
	131	100.0

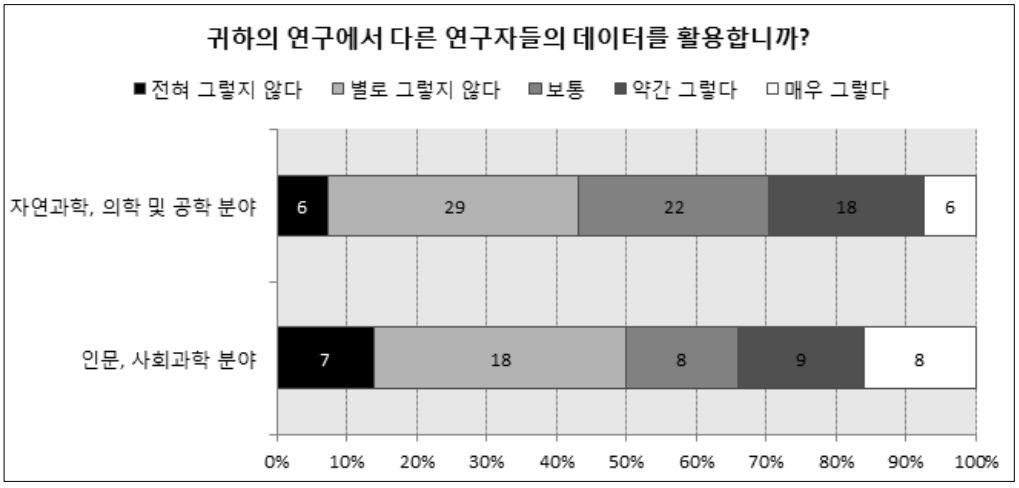
연구데이터를 공유하는데 있어 응답자들 간에 가장 보편적으로 활용되는 방식은 이메일이었고 이는 복수응답을 허용하여 수집된 256개의 응답 중 41.8%를 차지하였다(〈그림 8〉 참조). 이동식 매체는 31.3%의 응답자가 데이터 공유에서 활용한다고 응답하였고, 데이터 보관 장소에서도 언급되었던 웹 하드가 전체 응답의 16.8%를 차지하여 네트워크 저장 공간에서의 데이터 공유가 적지 않게 이루어짐을 알 수 있다. 인터넷 공간을 데이터 공유에 활용하는 또 다른 예로 접근 제한이 있는 웹사이트에 게시한다는 응답도 있었다.



〈그림 8〉 데이터 공유 방식

뿐만 아니라 소수이기는 하지만 국내 외 연구데이터 아카이브를 통해 데이터를 공유하는 연구자들이 있었다(〈그림 8〉 참조). 설문조사에서 연구데이터 아카이브의 정의를 '정부기관이나 연구비 지원기관, 학술지 등에서 연구데이터의 활용 및 보존을 위해 구축한 관리 시스템/DB'라고 제시하였으며, 응답자들이 활용하는 아카이브의 이름을 제시하도록 하였다. 아카이브를 통해 데이터를 공유한다는 응답자 8명 중 2명만이 여기에 응답하였는데 GenBank를 활용한다는 연구자와 함께 한국신약학회 홈페이지, 한국기독교학회 홈페이지, 학술연구정보센터, 한국연구재단 부설 KCI 사이트 등 다양한 웹 공간 및 서비스를 데이터 공유에 활용하는 연구자도 있었다.

데이터의 공유 현황과 더불어 다른 연구자들이 생산한 데이터를 재이용하는 정도를 살펴보면, 인



〈그림 9〉 데이터의 재이용 정도

문·사회과학 분야와 자연과학·의학 및 공학 분야 응답자들 모두 절반 혹은 절반에 가까운 비율로 타인의 데이터를 활용하지 않는다고 응답하였다 (〈그림 9〉 참조). 그러나 다른 연구자들의 데이터를 재이용한다는 응답이 두 집단 모두 30% 가량인 것으로 나타나 데이터를 공유하는 응답자들의 비율보다는 높은 것으로 나타났다.

타인의 데이터를 획득하는 방식으로는 출판된 논문에서 데이터를 추출하였다는 응답과 개인적으로 연구자들에게 연락하였다는 응답이 거의 비슷한 비율로 나타났다 (〈표 6〉 참조). 타인의 데이터를 수집하는데 있어 연구데이터 아카이브를 활용한다는 연구자들도 26명인 것으로 나타나 데이터 공유를 위해 아카이브를 활용하는 연구자들보다 많다는 것을 알 수 있다. 이들이 타인의 데이터 재이용을 위해 활용하는 아카이브의 예로는 기업정보를 제공하는 DART 전자공시시스템, 사회과학자료원이나 통계청, 한국학술정보(KISS)와 한국교육학술정보원(RISS), 기초학문자료센터, 세종언어자원, 한국신약학회 및 한국기독교학회 홈페이지, The Cooperative Association for Internet Data Analysis(CAIDA), NASA Public Database 등이 있었다.

〈표 6〉 타인의 데이터 획득 방식*

	빈도	퍼센트
출판된 논문의 본문에서 데이터추출	74	38.5
개인적으로 연구자에게 연락	73	38.0
국내 · 외 연구데이터 아카이브	26	13.5
연구자의 웹사이트	19	9.9
합계	192	100.0

*복수응답 허용

마지막으로 연구데이터의 관리 과정을 데이터의 생산/수집, 검색, 메타데이터 작성, 데이터의 분석, 단기간 보존 및 장기간 보존으로 구분하여 각각에 대한 만족도를 5점 척도로 조사하였다 (〈표 7〉 참조). 전체적인 만족도의 평균은 3.3점으로 보통 수준의 만족도를 보이고 있었으며, 가장 만족도가 낮은 데이터 관리 단계는 장기적인 데이터 보존과 메타데이터 작성이었다. 특히 메타데이터 작성의 경우 데이터의 신뢰성 확보와 보존에 필수적인 작업이지만 해당사항 없음에 표시한 응답자들이 다른 항목에 비해 많았다. 대학 내에서 생산되는 연구데이터를 체계적으로 관리하기 위해서는 장기적인 보존과 메타데이터 제공에 대한 지원이 요청되며 이를 실현할 수 있는 서비스 개발이 필요함을 알 수 있다.

<표 7> 데이터 관리 과정에 대한 만족도

	만족도평균
데이터의 분석 (n=130)	3.6
데이터의 생산/수집 (n=131)	3.5
과제 수행기간 동안(단기간) 데이터보존 (n=128)	3.4
데이터의 검색 (n=128)	3.4
과제수행 종료 이후(장기간) 데이터보존 (n=126)	3.0
데이터에 대한 설명 및 관리항목(메타데이터) 작성 (n=121)	2.9

Ⅴ. 결론 및 제언

본 연구의 설문 응답자들은 거의 모두 대학 교수이었으며 2010년과 2011년에 한국연구재단의 연구비 지원을 받은 연구자들이므로 활발히 연구를 수행할 수 있는 기반을 가지고 있다고 가정할 수 있다. 이들의 연구 분야를 크게 인문·사회과학 분야와 자연과학·의학 및 공학 분야로 구분하였으며 두 집단 간의 연구데이터 유형의 차이를 분석하였다. 가장 많은 수가 응답한 실험데이터의 경우 대다수가 이공, 의학 분야의 연구자들이 생산하는 것으로 나타났으며 계산데이터 및 관찰데이터도 비슷한 경향을 보였다. 그러나 두 번째로 많은 수가 응답한 설문데이터의 경우 대다수가 인문·사회과학 분야 연구자들의 응답이었으며, 면담데이터, 정부통계, 기록 역시 이 분야의 연구자들이 주로 생산, 수집하는 원 자료인 것으로 나타났다. 이를 바탕으로 학문 영역에 따라 생산, 수집되는 연구데이터의 유형에 차이가 있음을 알 수 있다.

연구데이터의 관리 방식과 관련하여 본 연구에서는 데이터의 가치를 높이는 처리 방식과 데이터 보관 장소 및 데이터의 보존 연한에 대한 인식을 살펴보았다. 또한 현재 응답자들이 속한 프로젝트나 실험실 또는 기관에서 존재하는 연구데이터의 관리체계를 조사하였다. 데이터의 처리 방식에 있어 다수의 연구자들이 논문 등의 관련 정보를 데이터에 연결시키거나 추가적으로 생산된 데이터를 병합한다고 응답하였다. 그러나 이렇게 연구수행 과정에서 자연스럽게 이루어질 수 있는 작업 이외에 데이터에 대한 기록화 작업, 예를 들어 주석을 작성한다든지 메타데이터를 제공하는 활동은 상대적으로 적은 수의 연구자들이 수행하고 있다고 응답하였다. 데이터 보관 장소의 대부분은 개인 PC와 이동식 매체인 것으로 나타났으며 그 이외에 실험장비나 소속기관 서버, 웹 저장 공간 등을 활용하는 것으로 나타났다. 이는 장기적인 관점에서 데이터의 손실에 대한 위험이 높은 방식들이므로 이에 대한 인식 제고와 대학에서 생산되는 연구데이터의 보존을 위한 전략이 논의될 필요가 있다.

데이터의 유용성이 유지되는 연한에 대한 인식에 있어서 78%의 응답자들이 10년 미만이라고 응

답하여 중단기적인 데이터 보존에 대한 요구가 높은 것으로 나타났다. 또한, 10년 이상의 보존 연한을 기대하는 응답자들은 자연과학·의학 및 공학 분야 연구자들의 비율이 인문·사회과학 연구자들보다 높은 것으로 조사되었다. 기존 연구에서는 세부적인 분야 별로 데이터의 유용성이 유지되는 기간에 대한 인식에 차이가 있었으므로 두 집단에 속한 학문 분야 별 특수성이 고려되어야 할 것이다. 뿐만 아니라 데이터의 유형 별로 보존기한에 대한 인식에 차이를 보일 수 있는 가능성이 있으므로 이에 대한 추가적인 논의가 필요할 것이다. 대다수 응답자들의 소속기관이나 프로젝트에서 연구데이터의 관리 체계를 갖추고 있지 못한 것으로 나타났으며, 관리체계가 존재한다고 응답한 연구자들의 경우 단기적인 데이터 관리 시스템과 데이터 관리 지침이나 내규, 데이터 관리에 대한 교육에 대한 응답이 많았다.

연구데이터 공유의 정도와 그 범위를 보았을 때 다른 이들이 응답자들의 데이터에 쉽게 접근할 수 있다고 응답한 비율이 20% 미만인 것으로 나타나 데이터의 공유 또는 공개의 정도는 낮은 것으로 나타났다. 데이터의 공유 범위를 살펴보면 연구자들과의 공유나 소속 연구팀을 중심으로 하는 데이터 공유가 일반적이었다. 데이터를 오픈 액세스로 공개한다는 응답자는 약 5%인 7명인 것으로 나타나 데이터를 누구나 활용할 수 있게 공개하는 정도는 낮은 것으로 나타났다. 데이터를 공유할 때 이메일과 이동식 매체가 가장 많이 활용되는 방식들이었고 상당수의 응답자들이 웹 하드와 같은 인터넷 저장 공간을 활용하고 있었다. 국내외의 연구데이터 아카이브를 활용하여 데이터를 공유한다는 응답은 소수 존재하였다.

타인의 연구데이터를 활용하는 데이터 재이용 정도를 살펴보았을 때 데이터 공유 정도에 비해 그 비율이 높은 것으로 나타났다. 약 30%의 응답자들이 타인의 연구데이터를 활용하고 있었고 대다수가 출판된 논문에서 데이터를 추출하거나 개인적으로 연구자에게 연락하여 데이터를 획득한다고 하였다. 연구데이터 아카이브를 이용하여 타인의 데이터에 접근한다는 응답자도 26명인 것으로 나타나 데이터 공유를 위해 아카이브를 이용하는 응답자보다 세 배 이상 많은 것으로 나타났다.

연구데이터 관리 과정에 대한 응답자들의 만족도를 조사한 결과 보통 수준의 만족도를 보였으며 데이터의 장기보존과 메타데이터 작성에 있어 낮은 만족도를 보였다. 이는 대다수의 응답자들이 연구데이터의 관리 체계가 존재하지 않는 환경에서 연구를 수행하고 있고, 장기적인 보존을 위한 시스템이나 데이터 관리를 위한 비용이 준비되지 않은 경우가 대부분인 현실을 반영하는 것이라고 할 수 있다. 메타데이터 작성을 통해 데이터에 대한 신뢰성과 접근성이 향상될 수 있고 이는 데이터의 공유와 재이용을 촉진시키는 요인이 될 수 있다. 또한, 장기적인 데이터의 보존은 전문적인 서비스의 개발과 이를 유지하는 비용의 문제로 인해 기관 또는 국가 차원의 지원이 필수적이다. 대학의 연구 활동에서 생산되는 데이터의 체계적 관리에 대한 필요성을 제고하고 이를 실현 가능하게 하는 서비스에 대한 연구와 노력이 지속적으로 이루어져야 할 것이다.

참고문헌

김은정. 연구데이터 수집에 영향을 미치는 요인 분석. 박사학위논문, 중앙대학교 대학원 기록관리학과 기록물관리학 전공, 2012.

신영란. 인문사회 연구데이터 아카이브의 발전방향에 관한 연구, 석사학위논문, 이화여자대학교 정책과학대학원 기록관리 전공, 2012.

American Council on Learned Societies (ACLS). *Our Cultural Commonwealth: the Final Report of the American Council of Learned Societies commission on Cyberinfrastructure for the Humanities and Social Sciences*. 2006. 〈www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf〉 [cited 2012. 8.18].

Beagrie, Neil, Robert Beagrie, and Ian Rowlands. "Research Data Preservation and Access: The Views of Researchers," *Ariadne*, Vol.60(2009). 〈http://www.ariadne.ac.uk/issue60/beagrie-et-al/〉 [cited 2012.8.18]

Borgman, Christine L., Jillian C. Wallis, and Noel Enyedy. "Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology." *Lecture Notes in Computer Science*, No.4172(2006), pp.170-183.

Borgman, Christine L.. "Data, Disciplines, and Scholarly Publishing." *Learned Publishing*, Vol.21(2008), pp.29-38.

Borgman, Christine L. "The Digital Future is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly*, Vol.3, No.4(2009). 〈http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html〉[cited 2012.8.18]

Carlson, Samuelle and Ben Anderson. "What are Data? The Many Kinds of Data and Their Implications for Data Re-Use," *Journal of Computer-Mediated Communication*, Vol.12, No.2(2007). 〈http://jcmc.indiana.edu/vol12/issue2/carlson.html〉 [cited 2012.8.18].

Faniel, Ixchel M. and Trond E. Jacobsen. "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data." *Computer Supported Cooperative Work*, Vol.19(2010), pp.355-375.

Griffith, Aaron. "The Publication of Research Data: Researcher Attitudes and Behaviour," *International Journal of Data Curation*, Vol.1, No.4(2009), pp.46-56.

Hedstrom, Margaret and Jinfang Niu. "Incentives for Data Producers to Create "Archive-Ready" Data: Implications for Archives and Records Management," *Proceedings of 2008 Society*

of American Archivist Research Forum. 2008. 〈http://www2.archivists.org/sites/all/files/M-HedstromJ-Niu-SAA-ResearchPaper-2008.pdf〉 [cited 2012.8.18.]

Higgins, Sarah. "The DCC curation lifecycle model," *International Journal of Digital Curation*, Vol.3, No.1(June 2008), pp.134-140.

Hey, Tony and Anne E. Trefethen. "Cyberinfrastructure and e-Science." *Science*, Vol.308(May 2005), pp.134-140.

Humphrey, Charles. e-Science and the life cycle of research, 2006. 〈http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc〉 [cited 2012.8.18].

Karasti, Helena, Karen S. Baker, and Eija Halkola. "Enriching the Notion of Data Curation In E-Science: Data Managing and Information Infrastructuring in the Long-Term Ecological Research (LTER) Network," *Computer Supported Cooperative Work*, Vol.15(2006), pp.321-358.

Lyon, Liz. "eBank UK: Building the Links between Research Data, Scholarly Communication and Learning," Ariadne, Vol.36, 2003. 〈http://www.ariadne.ac.uk/issue36/lyon/〉 [cited 2012.8.18]

National Science Foundation (NSF), *Revolutionizing Science and Engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*.(Arlington, VA: National Science Foundation, 2003).

Perry, Carol Marie. "Archiving of Publicly Funded Research Data: A Survey of Canadian Researchers," *Government Information Quarterly*, Vol.25(2008), pp.133-148.

Swan, Alma and Sheridan Brown. *To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network*, 2008. 〈http://eprints.soton.ac.uk/266742/〉 [cited 2012.8.18].

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. "Data Sharing by Scientists: Practices and Perceptions," *PLoS ONE*, Vol.6, No.6(June 2011). 〈http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101〉 [cited 2012.8.18].

Whyte, Angus, Dominic Job, Stephen Giles, and Stephen Lawrie. Meeting Curation Challenges in a Neuroimaging Group. *International Journal of Digital Curation*, Vol.3, No.1(2008), pp.171-181.

Zimmerman, Ann S. "New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data," *Science, Technology & Human Values*, Vol.33, No.5(2008), pp.631-652.