Data Citation Index를 기반으로 한 연구데이터 인용에 관한 연구

Study about Research Data Citation Based on DCI (Data Citation Index)

조 재 인 (Jane Cho)*

목 차

초 록

연구데이터의 개방과 공유는 연구의 효율성과 연구 과정의 투명성을 제고할 뿐 아니라, 데이터 통합과 재해석을 통해 새로운 과학으로의 창출도 가능하다. 서구를 중심으로 연구데이터 공개와 재사용을 위한 다양한 정책이 개발되면서 표준적인 인용 체계도 자리를 잡아가고 있다. 본 연구는 연구데이터 인용색인 DCI(Data Citation Index)를 기반으로 연구데이터의 구축 규모와 인용 정도를 파악하고, 기술통계분석과 Kruskal-Wallis H 분석을 통해서 고인용 데이터의 특성과 인용 경향을 분석해 보았다. 또한 알트매트릭스(Altmetrics) 분석 도구인 Impactstory를 통하여 연구데이터의 사회적 영향력도 진단해 보았다. 그 결과 연구데이터의 규모는 유전학과 생명공학 분야가 압도적으로 크지만, 다수 인용된 분야는 인구, 고용 등 경제・사회과학분야인 것으로 나타났으며, UK Data Archive, ICPSR(Inter-University Consortium For Political And Social Research)에 구축된 연구데이터가 가장 많이 인용되고 있는 것으로 분석되었다. 또한 데이터세트보다는 조사방법과 연구방법론이 포함된 데이터스터디가 높은 피인용도를 보이는 것으로 나타났으며, 연구데이터의 알트매트릭스 분석 결과에서도 사회과학분야의 데이터스터디가 상대적으로 많이 참조되고 있는 것으로 나타났다.

ABSTRACT

Sharing and reutilizing of research data could not only enhance efficiency and transparency of research process, but also create new science through data integrating and reinterpretationing. Diverse policies about research data sharing and reutilizing have been developing, along with extending of research evaluating spectrum that across research data citation rate to social impact of research output. This study analyzed the scale and citation number of research data which has not been analyzed before in korea through data citation index using Kruskal-Wallis H analysis. As result, genetics and biotechnology are identified as subject areas which have most huge number of research data, however the subject areas that have been highly cited are identified as economics and social study such as, demographic and employment. And Uk Data Archive, Inter-university Consortium for Political and Social Research are analyzed as data repositories which have most highly cited research data. And the data study which describes methodology of data survey, type and so on shows high citation rate than other data type. In the result of altmetrics of research data, data study of social science shows relatively high impact than other areas.

키워드: 연구데이터, Data Citation Index, ICPSR, 데이터 레포지토리
Research Data, Data Citation Index, ICPSR, Data Repository

* 인천대학교 문헌정보학과 부교수(chojane123@naver.com)
논문접수일자: 2016년 1월 22일 최초심사일자: 2016년 1월 26일 게재확정일자: 2016년 2월 11일
한국문헌정보학회지, 50(1): 189-207, 2016. [http://dx.doi.org/10.4275/KSLIS.2016.50.1.189]

1. 서 론

연구데이터의 개방과 공유는 연구의 효율성과 연구 과정의 투명성을 제고할 뿐 아니라 데이터 통합과 재해석을 통해 새로운 과학으로의 창출도 가능하다. 공공 기금으로 수행된 연구 성과의 부산물을 공개하여 재사용할 수 있도록 하는 다양한 정책이 수립되면서, 자연과학 분야에서부터 인문사회과학 분야에 이르기까지 원시데이터를 공개하고 공유하는 움직임이 전 세계적으로 확산되고 있다(Sayogo and Pardo 2013). 2013년 G8 과학 장관 회의에서는 연구데이터 공개를 통한 새로운 시대의 과학 협력이 천명되었으며(Department For Business, Innovation and Skills 2013), 백악관 과학기술정책국에서도 연구 개발 지출액이 연간 1억 달러를 넘는 정부 기관에 대하여 6개월 이내에 연방예산에 의한 연구 성과물 즉, 피어 리뷰 출판물과 기밀 연구 이외의 연구데이터를 공개하도록 하였다(The office of Science and Technology Policy 2013). 이에 따라 연구 지원 단체, 저널 출판사 등은 연구데이터 저장소 마련을 시작하고 있으며, RDA(Research Data Alliance, rd-alliance.org)와 같은 국제 단체도 등장해, 연구데이터의 공유와 재사용을 위한 운동을 지원하고 있다.

한편, 이러한 움직임과 더불어, 연구 성과 측정 방식에 있어도 기존과 같이 논문의 피인용도를 측정하는 방식에서 연구의 부속물인 데이터의 인용도를 포함하는 포괄적 체계로 확장 움직임을 보이고 있다(西薗 由依 2013). 특히 그동안 독보적인 인용색인을 제공해 온 톰슨로이터(Thomson Reuters)가 DCI(Data Citation Index)를 통해 연구데이터 피인용도에 대한 정보를 제공하기 시작하면서, 새로운 연구 영향력 평가 시대의 도래가 예고되고 있다.

본 연구는 DCI를 통해 연구데이터의 구축 규모와 주제 분야를 파악하고 고인용 데이터를 추출하여 그 특성과 인용 경향을 자세히 분석해 본다. 본 연구의 목적을 좀 더 상세히 기술하면 다음과 같다.

첫 번째, 연구데이터 공유 및 재활용의 필요성을 살펴보고 이를 위한 국제적 움직임을 조망해 본다. 또한 연구데이터 인용이 가지는 의미와 표준적 인용을 위한 기본 조건을 고찰해 본다. 두 번째, DCI를 기반으로 기술통계분석을 수행해 데이터 규모와 유형, 주제별 분포와 다수 생성 국가 등을 파악해 본다. 세 번째, 고인용 연구데이터 500건을 추출하여 그 특성과 인용 경향을 자세히 조사하고 Kruskal-Wallis H 분석을 통해 데이터 유형과 주제 분야가 피인용도에 어떠한 영향을 주고 있는지 분석해 본다. 네 번째, 연구데이터가 가지는 사회적 영향력을 알트매트릭스 측정 도구인 Impactstory를 통해서도 파악해 보며, 측정 결과를 DCI의 피인용 정도와 비교함으로써 연구데이터의 영향력을 다각도로 진단해 본다.

2. 이론적 배경

2.1 연구데이터의 필요성과 최근 동향

생물학 분야에서의 유전체 빅데이터는 차세대 염기서열분석 방법(Next Generation Sequencing)의 도입으로 기하 급수적으로 등장하고

있으며 공개 범위도 확대되고 있다. 최근 2년간 유전체 빅데이터의 생산량은 페타(Petabyte) 수준에 이르고 있으며, 매년 2배 이상 증가하고 있다. 이에 대용량 염기 서열 정보와 다양한 생물학 연구 분야의 정보가 통합되어 대사유전체학(Metabolomics)과 같은 생물학적인 현상을 설명할 수 있는 새로운 분야도 등장하고 있다(김운봉, 김용민, 양진옥 2014). 또한 그동안 생물학 논문의 실험 결과는 70% 이상이 재현할 수 없다고 지적되어 왔으나, 연구데이터의 공개와 재사용 경향에 의해 추가 시험이나 재현이 가능해져 연구의 투명성도 향상되고 있다(池內有爲 2014).

이와 같이 연구데이터 공유를 통해 과학 발전을 촉진시키고 연구의 투명성을 향상시키기 위해서는 데이터의 공개가 전제되어야 한다. 데이터 공개 필요성을 좀 더 구체적으로 정리해보면, 첫 번째, 연구자 커뮤니티에 있어서 잘 관리되어 공개된 데이터는 재사용 및 통합을 통해 새로운 과학으로 창출이 가능하다. 두 번째, 연구결과에 대한 반복적 실험과 연구 방법론의 검증과 발전이 가능해진다. 세 번째, 연구지원기관의 입장에서는 연구 결과의 검증을 통해 과학 발전을 도모할 수 있으며, 중복 연구 비용을 절감시킬 수 있다. 네 번째, 연구자 개인의 입장에서는 연구의 과학적 처리 과정을 추적할 수 있으며 연구자간 데이터 교환을 통해 협력의 기회도 증가시킬 수 있다.

연구데이터에 대한 OEDC 원칙과 가이드 라인에서는 공공 기금 지원을 받은 연구데이터는 공적 관심 중에 생산되었으므로 지적 소유권을 침해하지 않는 수준에서 시의적절하고 책임있게 공유되어야 한다고 천명하고 있다(OEDC 2007). 그에 따라 연구데이터 공개 의무화를 위한 다양한 제도적 장치가 마련되고 있다. 경제개발협력기구(OEDC)와 미국 백악관 과학기술정책국(OSTP), 유럽 의회 등에서는 공공 기금으로 수행된 연구의 원시 데이터 공개를 지시하고 있으며, 연구 보조금 신청시 연구데이터를 어떻게 저장하고 공유할 것인가를 기재한 '데이터 관리 계획(DMP: Data Management Plan)'의 제출을 의무화하는 경우도 증가하고 있다. 2013년 1월부터는 미국 과학 재단(National Science Foundation)의 보조금 신청 지침이 변경되어 연구 산출물의 명칭이 "출판물(Publications)"에서 "생산물(Products)"로 바뀌었으며, 그에 따라 논문과 함께 연구데이터가 제출되어야 한다(National Science Foundation 2012). 한편, 일본에서는 문부 과학성 및 후생 노동성 과학기술진흥기구의 지원을 받은 연구의 경우, 공모 단계에서부터 데이터 공유에 대한 협력이 권고되고 있으며(池內有爲 2014), 우리나라에서도 한국과학기술정보연구원(KISTI)에서 과학데이터 공유와 보존을 위한 시스템인 P-Cube (Platform For Convergence Research and Unification of Big E-Resources)를 운영하고 있다(김지현 2014). 또한 한국연구재단의 기초학문자료센터(Korea Research Memory, KRM)를 통해서는 인문사회분야 연구의 원 자료가 수집되고 있어, 연구데이터 공개에 대한 관심이 고조되고 있다.

한편, 분야를 넘어선 연구데이터 공유가 활성화되기 시작하면서, 연구데이터 동맹 RDA (Research Data Alliance)가 "장벽없는 데이터 공유"를 슬로건으로 발족되었다. 여기에서는 메타데이터 등의 기술 기반과 각종 정책 개

발을 포함한 사회 기반 구축 노력이 병행되고 있다. 또한 연구데이터에 영구적인 식별자인 DOI를 부여하고 데이터의 발견, 인용, 추적과 영향력 측정에 이바지하는 것을 목표로 DataCite (Https://www.datacite.org)가 창설되기도 하였다. 한편, 미국의 많은 대학에서도 연구자의 데이터 관리를 지원하기 위한 RDM(Research Data Management)서비스가 시작되고 있으며, 플랫폼으로 도서관의 기관 레포지토리, 데이터 포털, 전문 데이터센터, 주제별 레포지토리 등이 사용되고 있다. 더불어, 연구 결과를 해석하지 않고 데이터 생성 방법, 구조, 재사용 가능성 등을 제시하고 있는 데이터 저널의 창간도 이어지고 있다(池內有爲 2014).

2.2 연구데이터의 인용

타인의 데이터를 활용했을 때 기여자에게 적절한 크레딧이 돌아갈 수 있도록 하기 위해서는 데이터 인용이 반드시 필요하다. 데이터 인용은 기여자를 표시함으로써 공정한 보상이 돌아갈 수 있도록 할 뿐 아니라, 관련된 출판물과 연계시킴으로써 과학적 질문에 빠르게 접근할 수 있도록 한다(DataOne Education Modules Homepage).

데이터 인용에는 저자, 조사자, 데이터 생성자, 데이터가 공개된 일자, 데이터 소스, 제목, 버전, 데이터 형식, 아카이브와 디스트리뷰터 (Distributer) 등의 요소가 포함될 수 있으며, 그 밖에 고유식별자, 주제어, 에디터, 관련된 저작이 추가될 수 있다. 하버드 대학도서관에서는 데이터 인용의 요소로 저자명, 제목, 공개 혹은 배포날짜, 데이터가 공개된 레포지토리, 데이터의 판/버전/권, 분석에 사용된 소프트웨어, 고유 식별자 등의 접근 정보를 제시하고 있으며 (Havard Library Citing Your Data Homepage), DataCite(2015)가 정의한 Metadata Schema For The Publication and Citation of Research Data Version 3.1에서는 식별자, 저자, 타이틀, 출판사, 출판년을 필수 요소로, 주제, 기여자, 연도, 리소스타입, 관련식별자, 기술, 지역을 권장요소로 정의하고 있다. 인용시에는 'Creator (Publication year): Title. Publisher. Identifier'와 같은 규칙으로 기술하도록 제시하고 있으며, 예시하면 아래와 같다(Datacite Homepage).

- Geofon Operator (2009): Gefon Event Gfz2009Kciu (Nw Balkan Region). Geoforschungszentrum Potsdam (Gfz).Http://Dx.DOI.Org/10.1594/Gfz.Geofon.Gfz2009Kciu

한편, 연구자가 데이터를 출판하기 위해서는 영구식별자를 부여받아야 하며, 데이터를 출판하기로 결정한 데이터 퍼블리셔로부터 인용 정보를 획득해야 한다. 데이터를 출판할 수 있는 장소는 데이터 포털, 주제별 저장소, 전문 데이터 센터, 아카이브, 테마별 저장소, 기관 저장소 등이 있으며, 데이터 저널이나 프로젝트 기관의 웹 사이트를 통해 공개되거나 Figshare 같은 클라우드 기반의 자기 공개 시스템을 통해 출판되기도 한다.

영구식별자는 해당 연구데이터에 대한 고유 접근을 위해 반드시 필요한데, 가장 많이 쓰이고 있는 DOI 이외에, ARK, UUID 등이 존재한다. DOI는 Registration Agency로부터 부여된

번호 정보를 10.1234/NP5678, 10.5678/ISBN-0-7645-4889-4, 10.2224/2004-10-ISO-DOI와 같이 표시하는데 Datacite를 통해 부여받게 된다. 한편, 객체에 대한 장기 엑세스를 위해 제시되고 있는 ARK는 http://ark.cdlib.org/ark:/13030/tf5p30086k와 같은 방식으로 표시되며, 'Practically Unique Identifiers'인 UUID는 하이픈으로 구분된 5개의 그룹과 36개의 캐릭터로 구성되어, 550e8400-e29b-41d4-a716-446655440000와 같은 방식으로 표시된다(DataOne Education Modules Homepage).

2.3 Data Citation Index (DCI)

각종 연구 부산물이 증가하고 있으나 표준적인 인용 방식 부재로 인하여 재활용이 활성화되지 못하였으며, 그로 인해 영향력 측정도 어려웠다. 그러나 2012년 11월부터 톰슨 로이터가 Data Citation Index(DCI)를 통해 연구데이터의 색인과 인용정보를 제공함으로써, 데이터 인용에도 활기를 띠게 되었다.

톰슨로이터는 Web of Science에 인덱스(Index)할 학술지를 선정하는 절차와 마찬가지로 연구데이터를 인덱스할 데이터 레포지토리를 엄격한 기준으로 선정하고 있다. 데이터 레포지토리는 데이터 홀더(Data Holder)나 디스트리뷰터(Distributer)를 의미하는데, 연구비 지원 정보와 표준화된 인용 정보를 가지고 있는 영어 기반의 데이터 레포지토리를 그 대상으로 하고 있다. 대상 데이터 레포지토리 중 지속성과 안정성이 보장되고, 수록 범위의 양과 질, 저자의 다양성, 학술 커뮤니케이션 측면의 유용성, 데이터 큐레이션의 안정성이 보장되는 데이터 레포지토리가 선정되고 있다. 톰슨로이터는 전 세계적으로 1,000개 정도의 데이터 레포지토리를 발견하였는데, 그 중 150개 정도가 이러한 기준을 충족하고 있다고 말하였으며, 분야로는 생명공학 분야가 전체의 55%로 가장 큰 비중을 차지하고 있다고 하였다(Force and Robinson 2014). DCI는 선정된 데이터 레포지토리로부터 데이터를 인덱싱해 해당 데이터와 관련된 WOS(Web of Science)의 연구 논문을 연계하고 있으며, 피인용 횟수를 제시할 뿐 아니라, DataCite 방식의 표준화된 인용정보를 제공하고 있다. 또한 데이터 스키마 표준을 유지하기 위하여 데이터 레포지토리들과 협력하고 있으며, 텍스트마이닝이나 수작업을 통해 텍소노미와 시소러스를 부여하고 있다.

2.4 선행연구

Force와 Robinson(2014)은 그동안 데이터 인용 정보는 학술논문에 비정형적인 방법으로 포함되어 있는 경우가 대다수였기 때문에 추적하기 어려웠다고 지적하였다. 그러나 표준적 방식의 데이터 인용이 증가하고 있어, 미래의 연구 영향력 평가 체계는 데이터 인용 매트릭스를 통합한 기존 방식의 확장된 형태가 될 것이라고 언급하였다. 또한 DCI는 정형적이건 비정형적이건 간에 데이터 인용에 대한 정보를 추출하여 공식화된 표준 인용 형식으로 바꾸고, 연구 논문과 데이터를 직접 연계시켜 줌으로써 연구데이터의 공유 활성화를 촉진하고 있다고 말하였다. 한편, Force와 Auld(2014)는 기여자에 대한 적절한 크레딧 부여 측면에서 DCI가 가지는 의미에 대하여 논하였다. 또한 이를 통

해 단일 창구로 다양하고 영향력 있는 데이터 레포지토리에 접근해 질 높은 데이터를 발견할 수 있을 뿐 아니라, 데이터와 관련된 연구 프로젝트를 이해할 수 있게 된다고 부연하였다.

Torres-Salinas, Martín-Martín, Fuente-Gutiérrez(2014)는 2013년 4월을 기준으로 DCI에 구축된 데이터 현황을 다음과 같이 분석하였다. 생명과학분야가 80%, 사회과학이 18%, 인문예술 2%, 기술공학 0.01%를 차지하고 있으며, 가장 큰 비중을 차지하고 있는 데이터 레포지토리로 생명과학분야의 Gene Expression Omnibus를 꼽았다.

국내에서는 아직까지 연구데이터에 대한 관심이 높지 않아, 비록 WOS를 구독하고 있는 경우라도 DCI까지 도입한 기관은 부재하다(Thomson Reuters 2015). 따라서 DCI에 접근하여 연구데이터의 규모와 현황, 고인용 데이터의 특징 등을 분석한 연구는 존재하지 않는다.

3. 연구의 방법

본 연구는 톰슨 로이터의 DCI를 기반으로 데이터를 수집하고, SPSS Statistics 21을 통해 기술통계분석과 Kruskal-Wallis H 분석을 수행하여 다음과 같이 해석하였다. 또한 피인용도에는 나타나지 않는 연구데이터의 영향력을 추적하기 위하여 알트매트릭스 측정 도구인 Impactstory(Impactstory.org)를 활용하여 분석을 수행하고, 그 결과를 DCI 분석 결과와 비교하였다. 조금 더 구체적으로 설명하면 다음과 같다.

첫 번째, 2006년도부터 2015년 3월까지 10년간 구축된 연구데이터를 추출하여 전반적인 데이터의 규모와 유형, 주제별 분포를 살펴보고 다수 생성 국가와 주요 데이터 레포지토리를 파악하였다.

두 번째, 피인용도가 높은 연구데이터 500건을 추출하여 고인용 데이터의 특성을 파악하였다. 기술 분석을 통해 주제 분야와 주제어를 분석해 보며, 데이터 레포지토리, 구축 연도, 데이터 유형, 데이터 조사 방법 등을 파악하였다.

세 번째, 데이터 유형과 주제가 인용도에 어떠한 영향을 미치는 지 파악하기 위하여 고인용 데이터 500건을 대상으로 기술통계분석과 세집단의 평균 차이를 검증하는 비모수기법인 Kruskal-Wallis H 분석을 수행하였다.

네 번째, 고인용 데이터 중 DOI가 있는 데이터 161건을 추출하여 오픈소스인 Impactstory를 통해 연구데이터의 알트매트릭스를 분석함으로써, 고인용 데이터가 가지는 사회적 영향력을 살펴보았고, DCI를 통해 도출된 피인용도와 차이가 나타나는지 비교해 보았다.

4. 분석 결과

4.1 연구데이터 공유 규모와 현황

먼저 DCI를 통하여 2006년도부터 2015년 3월까지 생성된 연구데이터의 규모를 분야별로 파악해 보았다. 최소 레코드 임계값을 100건으로 설정한 후 분석을 수행한 결과, 총 79개 분야의 3,379,301건의 데이터가 축적되어 있는 것으로 나타났다. 그 중 유전학(Genetics Heredity)이 1,772,377건으로 가장 많았으며, 생화학/분자

생물학(Biochemistry Molecular Biology)이 1,355,128건으로 두 번째로 많은 것으로 나타났다. 그 밖에도 지리(Geography), 해양학(Oceanography) 분야 등이 많은 데이터를 가지고 있었으며, 사회학(Sociology), 정치학(Political Science) 등의 사회과학분야도 눈에 띈다. 20순위까지 표시한 〈표 1〉에서는 나타나지 않았으나 정보학/문헌정보학 분야도 총 87건의 연구데이터가 누적되어 있는 것으로 집계되었다.

〈표 1〉 DCI의 분야별 연구데이터 규모 (2006-2015)

Web of Science 주제 범주	데이터건수
Genetics Heredity	1,772,377
Biochemistry Molecular Biology	1,355,128
Multidisciplinary Sciences	502,231
Geography	395,378
Geosciences Multidisciplinary	287,937
Oceanography	99,736
Crystallography	76,183
Plant Sciences	65,836
Chemistry Medicinal	50,508
Spectroscopy	40,861
Geochemistry Geophysics	32,903
Ecology	28,611
Marine Freshwater Biology	27,902
Water Resources	16,041
Sociology	13,563
Social Sciences Interdisciplinary	12,692
Cell Biology	10,778
Political Science	10,293
Economics	10,199
Demography	9,728

DCI에서는 데이터 형식을 레포지토리, 데이터스터디, 데이터세트의 세 가지의 유형으로 구분하고 있다. 첫 번째 레포지토리는 데이터 그 자체뿐 아니라, 조사방법론, 연구방법론과 같은 데이터에 대한 기술과 검색 메카니즘까지 포함하고 있는 포괄적 객체를 의미한다. 두 번째 데이터스터디는 연구에 사용된 데이터를 기술하고 있는 단위로, 데이터의 조사방법과 연구방법론, 데이터의 유형 등이 기술되어 있다. 마지막으로 데이터세트는 연구와 실험 산출물의 일부로서 데이터 그 자체를 의미한다. 따라서 데이터스터디와 같이 데이터 조사방법론 등을 설명하지 않으며, 레포지토리와 같이 검색 메카니즘을 포함하지도 않는다(Force and Robinson 2014). 분석 대상 연구데이터 중에서는 〈표 2〉와 같이 데이터세트가 3,200,752건으로 압도적으로 많은 것으로 나타났으며, 그 다음이 178,458건인 데이터스터디로 나타났다. 데이터 레포지토리는 단지 93건만이 존재하는 것으로 나타났다.

〈표 2〉 연구데이터 유형(2006-2015)

문서 유형	데이터건수
Data Set	3,200,752
Data Study	178,458
Repository	93

한편, 〈표 3〉과 같이 데이터를 공개하고 있는 국가는 미국이 압도적으로 많은 618,440건으로 나타났으며, 캐나다, 중국, 일본, 독일, 영국 순으로 많은 연구데이터를 구축하고 있는 것으로 나타났다. 우리나라는 15번째로 7,228건의 데이터가 존재하는 것으로 나타났으며 인덱스된 데이터는 대다수가 KISTI의 연구데이터 레포지토리에 축적되어 있는 것으로 파악되었다.

〈표 3〉 연구데이터 공유 국가(2006-2015)

순위	국가	데이터건수
1	Usa	618,440
2	Canada	104,117
3	China	85,998
4	Japan	54,181
5	Germany	46,488
6	United Kingdom	34,972
7	Italy	28,788
8	Netherlands	27,313
9	France	25,708
10	Switzerland	18,824
11	Australia	15,210
12	Spain	14,667
13	Sweden	12,257
14	Singapore	8,621
15	South Korea	7,228

DCI가 색인을 추출하는 데이터 레포지토리는 연구데이터가 출판되는 곳으로, WOS에서 학술저널과 비슷한 개념이라고 말할 수 있겠다. 다시 말해 학술논문이 출판되는 곳이 학술저널이라면 연구데이터가 출판되는 곳이 데이터 레포지토리라고 말할 수 있겠다. 앞에서 언급한 바와 같이 톰슨로이터는 엄격한 평가 기준을 가지고 인덱스할 데이터 레포지토리를 선정하고 있는데, DCI 분석 결과 가장 많은 양의 연구데이터를 가지고 있는 데이터 레포지토리는 Torres-Salinas, Martín-Martín, Fuente-Gutiérrez(2014)의 연구 결과에서와 같이 Gene Expression Omnibus로 분석되었다. 〈표 2〉에서 보여지는 바와 같이, Gene Expression Omnibus는 유전학 분야의 데이터 레포지토리로 807,009건의 연구데이터가 구축되어있는 것으로 조사되었다. 두 번째로 많은 양의 연구데이터를 포함하고 있는 데이터 레포지토리는 Figshare로 487,322건의 연구데이터를 가지고 있는 것으로 나타났는데, 이는 자기 공개가 가능한 클라우드 방식의 데이터 레포지토리로 다양한 분야의 연구데이터를 망라하고 있다.

〈표 4〉 주요 데이터 레포지토리(2006-2015)

데이터 레포지토리명	데이터건수
Gene Expression Omnibus	807,009
Figshare	487,322
U S Census Bureau Tiger Line Shapefiles	394,151
Uniprot Knowledgebase	337,653
Pangaea	271,094
Yeast Resource Center Public Image Repository	108,264
Arrayexpress Archive	89,509
Deg a Database of Essential Genes	88,360
Crystallography Open Database	73,368
Plant Transcription Factor Database	65,536
Worldwide Protein Data Bank	61,498
Sioexplorer	60,195
Human Metabolome Database	50,504
European Nucleotide Archive	48,380
Aspergillus Genome Database	47,797
Massbank	34,264
Emage Gene Expression Database	28,821
Partnership for Interdisciplinary Studies of Coastal Oceans Pisco	27,630
Mirbase	24,734
Lter Network Information System Repository	19,342
Candida Genome Database	18,068
Inter University Consortium for Political and Social Research	16,906

4.2 고인용 연구데이터 분석

4.2.1 고인용 연구데이터의 기술 분석 결과

피인용도가 높은 데이터를 추출하여 특징을

파악하기 위하여 DCI에서 상위 인용도를 보이는 500건의 연구데이터를 추출하였다. 추출된 데이터의 주제, 타입, 유형, 조사 방법론을 분석하고 어떠한 데이터 레포지토리에 출판된 연구데이터가 가장 높은 인용도를 보였는지 기술통계분석을 수행해 보았다.

첫 번째로 상위 인용도를 보인 500건의 연구데이터를 대상으로 주제 분야를 분석한 결과를 살펴보면 〈표 5〉와 같다. DCI에서는 WOS의 주제 카테고리를 이용하여 연구데이터를 그룹핑하고 있는데, 절대적인 연구데이터의 수는 앞서 살펴본 바와 같이 유전학과 생화학 분야가 가장 많았지만, 상위 인용된 연구데이터는 경제(Economics, 126건), 사회(Sociology, 101건), 인구(Demography, 75건), 건강관리/정책(Health Care & Policy, 68건) 분야 순으로 나타났다. 저자가 연구데이터에 부여한 주제어를 모두 추출하여 빈도분석을 수행한 결과도 역시 비슷하다. 〈표 6〉에서 제시하고 있는 것과 같이 500개의 연구데이터에서 총 20,226개의 주제어가 추출되었는데, 상위 빈도로 출현한 키워드는 가구(Households, 334회), 피고용인(Employees, 236회) 이외에도 교육배경(Educational Background, 225회), 정규직원(Full-Time Employment, 208회), 고용(Employment, 201회), 성별(Gender, 197회)순으로 나타나, 유전학 등 생명공학분야보다 인구, 고용 등 사회, 경제 분야 데이터의 인용도가 높은 것으로 나타났다. 다시 말해, 연구데이터의 절대적 구축량은 유전학, 생화학 분야가 많지만, 재활용성이 높아 다수의 후속 연구자에 의해 인용되고 있는 분야는 사회, 경제 분야인 것으로 분석되었다.

〈표 5〉 고인용된 연구데이터의 주제분야 (상위 500건)

주제분야	데이터건수	비중 %
Economics	126	25.2
Sociology	101	20.2
Demography	75	14
Health Care & Policy	68	13.6
Family Studies	55	11
Genetics & Heredity	13	2.6
Social Work	10	2
Education	8	1.6
Political Science	8	1.6
Ethnic Studies	7	1.4
Meteorology & Atmospheric Sciences	7	1.4
Criminology & Penology	5	1
Astronomy & Astrophysics	3	0.6
Ecology	3	0.6
Gerontology	3	0.6
Women'S Studies	3	0.6
Area Studies	2	0.4
International Relations	2	0.4
Marine & Freshwater Biology	1	0.2
Psychiatry	1	0.2
Psychology; Sociology; Health Policy & Services	1	0.2
Public Administration	1	0.2
Religion	1	0.2

〈표 6〉 고인용된 연구데이터에서 추출된 고출현빈도 주제어(상위 500건)

주제어	출현횟수
Households	334
Employees	236
Educational Background	225
Full-Time Employment	208
Employment	201
Gender	197
Economic Activity	194
Examinations	189
Educational Institutions	181

주제어	출현횟수
Employment Programmes	176
Employment History	174
Educational Grants	173
Conditions of Employment	168
Health	168
Furnished Accommodation	162
Absenteeism	154
Employment Services	151
Elderly	148
Disabled Persons	146
Domestic Responsibilities	146

두 번째로 데이터 유형을 살펴 본 결과, 데이터스터디가 고인용된 500건의 연구데이터 중 96%를 차지하고 있는 것으로 나타났다. 〈표 2〉와 같이 최근 10년내 구축된 데이터 전체를 대상으로 데이터 유형을 분석한 결과에서는 데이터세트가 가장 높은 비중을 차지하는 것으로 나타났지만, 고인용된 데이터 유형은 대부분 데이터스터디인 것으로 분석되었다. 연구데이터의 과학적 해석을 위해서는 데이터에 대한 충분한 설명이 필요하지만 아직까지 그렇지 않은 경우가 다수 존재한다. 따라서 데이터세트보다는 데이터의 조사방법과 연구방법론, 변수 등이 설명된 데이터스터디가 더 높은 인용도를 보이고 있다고 해석해 볼 수 있겠다.

〈표 7〉 고인용된 연구데이터의 유형

유형		빈도	비중 %
유효	Data Set	15	3.0
	Data Study	480	96.0
	Repository	5	1.0

세 번째로 고인용 연구데이터의 데이터 형식과 조사방법론을 살펴보았다. 분석 대상 데이터 중에는 하나의 레코드에 다양한 형식과 방법론이 혼재되어 있는 경우가 존재하였는데, 우선 기술된 데이터 형식을 추출하여 레코드별로 분석한 결과, 서베이 데이터가 118회로 가장 많은 것으로 나타났으며 그 밖에 게놈 유전자데이터, 추적데이터, 클리니칼 데이터 등 다양한 유형이 존재하는 것으로 나타났다. 한편, 조사방법론도 하나의 레코드에 다양한 방식이 혼합되어 있어, 이 역시 우선 기술된 방법론을 추출하여 분석하였는데, 인터뷰 방식과 설문 방식이 가장 많은 것으로 나타났으며, 그 밖에 랜덤 샘플링, 팀평가, 병원기록, 통계기록 등 다양한 조사방법론이 존재하는 것으로 나타났다.

네 번째로 연도별 연구데이터의 피인용도 평균을 살펴본 결과, 최신 년도로 올라갈수록 점차적으로 증가하다가, 2010년도부터 급증하는 추세를 나타냈으며, 2014년과 2011년에 출판된 연구데이터의 피인용도 평균이 가장 높은 것으로 나타났다.

다섯 번째로 고인용 데이터가 출판된 데이터 레포지토리를 분석해 본 결과, 인용순위 500위 내에 포함되는 데이터 레포지토리의 수는 단 20개로 요약되었다. Uk Data Archive와 Inter-University Consortium For Political and Social Research(ICPSR)가 각각 53.6%와 31.4%를 차지하여 전체의 85%가 이 두 개의 데이터 레포지토리에 포함되는 것으로 집계되었다. UK Data Archive는 영국에서 가장 큰 컬렉션을 가진 데이터 레포지토리로 인문사회 분야를 포괄하고 있으며, 미시간 대학교에서 운영하고 있는 ICPSR은 700개의 대학, 연구소가 컨소시엄으로 운영되는 사회과학 분야의 데이터 아카이브

〈표 8〉 고인용된 연구데이터 형식과 조사방법

데이터 형식	횟수	비중 %
Survey Data	118	41.70
Numeric Data	96	33.92
Administrative Data	20	7.07
Clinical Data	14	4.95
Transaction Data	8	2.83
Census	6	2.12
Textual Data	4	1.41
Transaction Data	3	1.06
Gene/Genome	3	1.06
Longitudinal Data	3	1.06
Observation Data	3	1.06
Fits Image	3	1.06
Census	2	0.71

조사방법	횟수	비중 %
Interview(Computer-Assisted Self Interview, Face-To-Face Interview, Telephone)	316	77
Questionnaire(Mail, Telephpne, On-Site, Self-Enumerated)	49	12
Random Sample	20	5
기타(Team Completes Assessment, Hospital Medical Records)	14	3
Compilation or Synthesis of Existing Material	9	2

〈표 9〉 고인용된 연구데이터 출판년도

연도	피인용도 평균	건수
2006	41.02	38
2007	64.37	29
2008	68.20	65
2009	66.10	174
2010	99.02	160
2011	126.84	13
2012	29.28	7
2013	101.15	13
2014	246.00	1

로 교육, 노령화, 범죄, 약물남용, 테러리즘 등 16개 분야의 특성화된 사회과학 분야 데이터 컬렉션을 포함하고 있다. 한편, 데이터 구축량이 많았던 Gene Expression Omnibus와 Figshare가 출판한 연구데이터에 대한 인용 정도는 높지 않아 20개의 데이터 레포지토리내에 포함되지 못하는 것으로 나타났다. Gene Expression Omnibus는 마이크로 어레이와 게노믹 데이터에 대한 퍼블릭 레포지토리로 데이터 구축량은 방대하지만, 사회과학분야와 같이 광범위하게

〈표 10〉 고인용 데이터가 출판된 데이터 레포지토리

데이터 레포지토리	데이터건수	비중 %
UK Data Archive	268	53.6
Inter-University Consortium for Political and Soci	157	31.4
Australian Data Archive	19	3.8
Manitoba Centre For Health Policy Population Healt	15	3
National Snow & Ice Data Center	7	1.4
Behavioral Risk Factor Surveillance System (Brfss)	6	1.2
Cancer Models Database	5	1
European Nucleotide Archive	5	1
Sloan Digital Sky Survey	3	0.6
Australian Antarctic Data Centre	2	0.4
Global Trade, Assistance, and Production: The Gtap	2	0.4
Oak Ridge National Laboratory Distributed Active A	2	0.4
Share - Survey of Health, Ageing and Retirement in	2	0.4
1000 Genomes - A Deep Catalog of Human Genetic Var	1	0.2
45 and Up Study	1	0.2
Born in Bradford Cohort Study	1	0.2
Broad-Novartis Cancer Cell Line Encyclopedia	1	0.2
Drosophila Genetic Reference Panel 2	1	0.2
Healthcare Cost and Utilization Project (Hcup)	1	0.2
International Food Policy Research Institute	1	0.2

재활용되거나 인용되지는 않는 것으로 나타났다. Figshare는 비록 다양한 분야를 포괄하는 레포지토리이지만 자기 기반의 클라우드형 레포지토리로 구축 데이터를 기술하는 메타데이터가 부족할 뿐 아니라, 표준화되지 않아 재활용성은 상대적으로 떨어지는 것으로 유추된다.

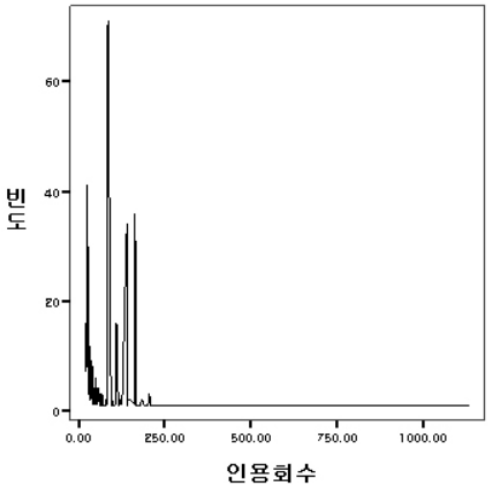
4.2.2 피인용 경향과 유형, 주제와의 관계 분석

여기에서는 고인용 연구데이터의 피인용 경향을 살펴보고, 데이터 유형과 주제가 과연 연구데이터의 인용에 어떠한 영향을 미치는 지 분석해 보았다.

(1) 고인용 연구데이터의 피인용 경향

고인용 연구데이터의 피인용 경향 분석 결과를 기술하면 다음과 같다. 가장 많이 인용된 연구데이터는 ICPSR에 저장된 "National Longitudinal Study of Adolescent Health(Add Health), 1994-2008"으로 총 1,135회 인용된 것으로 나타났다. Web of Science 핵심 컬렉션에서 889회, Biosis Citation Index에서 185회 인용된 것으로 나타났으며, 주제는 건강관리 및 서비스(Health Care Sciences & Services) 분야, 유형은 데이터스터디인 것으로 분석되었다. 510회

인용된 2번째 고인용 데이터는 "General Social Surveys, 1972-2006"으로 Web of Science 핵심 컬렉션 444회, Biosis Citation Index에서 3회 인용되었다. 이 연구데이터는 사회학 분야로 ICPSR에 출판된 것으로 나타났으며, 역시 유형은 데이터스터디인 것으로 분석되었다. 연구데이터의 인용 회수를 몇 개 구간으로 나누어 자세히 살펴 본 결과, 500회 이상 인용된 데이터는 단 2건뿐이며, 20-50회 사이에 절반 가까이가 분포되어 있는 것으로 나타났다. 백분위수로 구분해 보아도 아래와 같이 50회 정도에 50%의 연구데이터가 분포하는 것으로 나타났으며, 10%의 연구데이터만이 인용횟수가 164건 이상인 것으로 나타나, 〈그림 1〉의 그래프와 같이 긴꼬리 모양의 그래프를 보이고 있는 것으로 나타났다. 이러한 분석 결과를 가지고 유추해 볼 때, 전체 데이터를 대상으로 인용 경향을 분석한다면, 더욱 극단적인 긴 꼬리 모양을 보일 것으로 추론된다. 다시 말해 아주 극소수의 데이터만이 다수의 인용도를 보이고 있는 것으로 해석할 수 있겠다.



〈그림 1〉 백분위수를 기준으로 한 그래프

(2) 데이터 유형, 주제와 피인용도간의 관계 분석

앞서 말한 바와 같이 데이터 유형과 주제 분야가 피인용도에 어떠한 영향을 미치는지 파악하기 위하여 각각 기술통계분석과 Kruskal-Wallis H 분석을 실시한 결과는 〈표 12〉와 같이 나타났다.

〈표 11〉 인용횟수 구간별 데이터 수 및 인용횟수를 기준으로 백분위수

구간	데이터건수	비중 %	백분위수	인용횟수
20-50	238	47.6	10	24
51-100	129	25.8	20	26
101-500	131	26.2	25	27
501-1000	1	0.2	30	30
1000-	1	0.2	40	38
			50	53
			60	85
			70	85
			75	110
			80	140
			90	164

첫 번째, 데이터의 유형과 인용도간에 어떠한 차이가 나타나는지를 살펴보기 위하여 데이터유형을 독립변수로 설정하고 피인용도를 종속변수로 설정하여 Kruskal-Wallis H을 실시한 결과는 다음과 같이 나타났다. 먼저 기술통계분석을 통한 변수집단간의 차이를 살펴보면 세 집단의 평균 피인용도는 데이터 레포지토리가 압도적으로 높게 나타난 것을 확인할 수 있다. 〈표 7〉에서 분석한 고인용 연구데이터의 유형에서는 데이터스터디가 480건으로 96%를 차지하고 데이터세트가 15건으로 3%, 레포지토리가 5건으로 1% 분포하고 있는 것으로 나타났으나, 개별 연구데이터의 피인용건수 평균을 비교해 본 결과에서는 레포지토리가 116회로 가장 높게 나타났고 그 다음이 데이터스터디로 평균 78회, 마지막으로 데이터세트가 평균 37회로 가장 적게 인용되는 것으로 나타났다. 레포지토리는 500순위 내에 5개밖에 존재하지 않았지만, 그 자체에 데이터 구조와 조직을 반영하고 있으며, 검색 메카니즘까지 포함하고 있어 다른 데이터 유형에 비해 인용도가 높게 나타났을 것으로 추정된다. 데이터스터디도 데이터의 조사방법과 연구방법론, 데이터의 유형 등이 기술되어 있어 해석과 재활용이 용이해 자주 이용되는 반면, 데이터세트는 변수 및 실험 방법 등이 명확하게 제시되어 있지 않은 경우가 다수 존재해, 재활용성이 상대적으로 떨어지는 것으로 유추 가능하겠다. 세 집단의 차이에 대한 유의확률은 0.031로 나타나 집단간의 차이가 통계적으로 유의한 것으로 나타났다.

두 번째, 주제 분야와 피인용도 간에도 어떠한 관계가 존재하는지 파악하기 위하여 4개의 주제 분야를 독립변수, 피인용도를 종속변수로 설정해 Kruskal-Wallis H 분석을 실시하였다. 먼저 기술통계분석을 통한 변수집단간의 차이를 살펴보면 사회과학분야의 인용도가 평균 80회로 가장 높게 나타났고 그 다음이 71회로 생명공학 분야, 그 다음이 65회로 인문학 분야, 마지막이 34회인 자연/응용과학 분야 순인 것으로 나타났다. 구축된 데이터는 생명공학분야가 가장 많았지만 광범위한 재활용과 인용이 이루어

〈표 12〉 데이터 유형에 따른 평균 인용도 차이에 대한 기술통계분석과 Kruskal-Wallis H 분석결과

데이터타입	N	평균	평균에 대한 95% 신뢰구간		최소값	최대값
			하한값	상한값		
Dataset	15	37.3333	23.5846	51.0821	20.00	121.00
Datastudy	480	78.0771	71.2515	84.9027	20.00	1,135.00
Data Repository	5	116.6000	-47.4068	280.6068	20.00	344.00
합계	500	77.2400	70.5593	83.9207	20.00	1,135.00

	데이터타입	N	평균순위	인용회수	
인용회수	Dataset	15	154.60	카이제곱	6.949
	Datastudy	480	253.25	자유도	2
	Data Repository	5	274.30	근사 유의확률	.031
	합계	500			

〈표 13〉 데이터 유형에 따른 평균 인용도 차이에 대한 기술통계분석과 Kruskal-Wallis H 분석 결과

대주제	평균	N	합계
생명공학	71.0122	82	5,823.00
자연/응용과학	34.0769	13	443.00
사회과학	80.2879	396	31,794.00
인문학	65.1250	8	521.00
합계	77.3166	499	38,581.00

	데이터타입	N	평균순위	인용회수	
인용회수	생명공학	82	207.12	카이제곱	21.743
	자연/응용과학	13	118.88	자유도	3
	사회과학	396	263.57	근사유의확률	0.000
	인문학	8	231.13		
	합계	499			

지고 있는 분야는 〈표 5〉의 분석과 같이 사회과학분야인 것으로 나타났다. 주제 분야가 인용도에 영향을 주는지 통계적으로 파악하기 위하여 Kruskal-Wallis H 분석을 실시한 결과, 유의확률이 0.000으로 나타나 주제에 따라 피인용도에 통계적으로 유의한 차이를 보이는 것으로 나타났다.

4.2.3 연구데이터의 사회적 영향력

연구데이터의 인용은 데이터 기여자의 크레딧에 대한 인식 부족, 데이터에 대한 표준적 기술방식 미비 등의 요인으로 학술 논문에 비해 활성화되고 있지 않으며 인용 방식의 비정형화, 비표준화에 의해 발견이 쉽지 않은 것도 사실이다. 따라서 후속 연구에 의해 피인용되지는 않았으나, 연구데이터가 얼마나 연구자들에 의해 관심을 받고 있으며, 사회적으로는 어떠한 영향력을 가지고 있는지 다면적으로 살펴볼 필요가 있겠다.

최근에는 학술연구 영향력 평가에 있어 기존 피인용도 방식이외에 소셜미디어, 언론보도, 참고문헌관리도구 등으로부터 연구 성과가 언급된 빈도를 계산해 영향력을 측정하는 알트매트릭스(Altmetrics) 방식이 병용되고 있다. 아직까지 전통적인 연구 평가 방식을 대체하지는 못하였으나, 알트매트릭스는 기존 비블리오 매트릭스에 보완적 수단으로서 활용되고 있다. 더불어 다양한 연구(Mohammadi 2014; Zahedi et al. 2014; Haustein et al. 2015)에서 다각도로 검증이 이루어지고 있어 관심이 집중되고 있다. 알트매트릭스는 DOI와 같은 고유 식별자가 있는 경우, 연구 부속물의 영향력까지도 기계적으로 측정할 수 있는데, 본 장에서는 피인용도 매트릭스에서 보여지지 않았던 연구데이터의 사회적 영향력을 파악해 보기 위하여 오픈소스 알트매트릭스 분석 도구인 Impactstory를 통하여 측정해 보았다. 500건의 고인용 데이터 중 DOI가 존재하는 161건의 데이터를 대상으로 알트매트릭스를 측정한 결과는 다음과 같이 나타났다.

첫 번째, DOI가 존재하는 161건의 데이터에 대한 알트매트릭스 분석 결과, 페이스북, 블로그 등의 SNS 지표에서는 평가결과가 나타나지 않았으나, 참고문헌 관리도구인 멘델리(Mendeley, www.mendeley.com) 세이브드에서는 132건의 기록이 존재하는 것으로 나타났다. 그러나 단지 24개의 연구데이터에서만 기록이 존재해, 학술논문에 비하여 연구데이터의 알트매트릭스 민감도는 아직까지 매우 저조한 것으로 판단되었다. 가장 많이 멘델리에 저장된 연구데이터는 사회과학 분야의 서베이 데이터인 "Project on Human Development in Chicago Neighborhoods: Community Survey, 1994-1995"로 23회 저장되어 있는 것으로 나타났는데, 이는 DCI에서도 43회 인용된 것으로 나타난 데이터이다.

두 번째, 멘델리에 저장된 모든 연구데이터는 데이터스터디인 것으로 나타났다. 데이터세트와 레포지토리는 멘델리에 한 건도 저장되지 않은 것으로 나타났다. 구축량 자체가 많지 않은 레포지토리를 배제하고 설명하면, 데이터 피인용도 분석에서와 마찬가지로 데이터세트보다 데이터스터디가 많이 참조되고 있는 것으로 추정해 볼 수 있겠다.

세 번째, 주제가 부여되어 있는 데이터를 대상으로 분야를 살펴보면 사회과학분야(6.38회)가 가장 높고 그 다음 생명공학 분야(4.60회)가 높은 것으로 나타났다. 이는 DCI의 분야별 인용 결과와 유사한 경향을 보인다.

정리하자면 아직까지 연구데이터의 알트매트릭스 민감도는 높지 않지만, 참고문헌 관리도구인 멘델리에 남겨진 흔적을 통해 추적해 볼 수 있었다. 그 결과 사회과학 분야의 데이터스터디가 가장 높은 알트매트릭스 수치를 보여, 앞서 분석한 피인용 정도와 비슷한 경향을 보이는 것으로 분석되었다.

〈표 14〉 데이터 주제분야에 따른 알트매트릭스 평균

대주제	평균(Saved: Mendeley)	N
생명공학	4.60	5
자연/응용과학	1.00	1
사회과학	6.38	16
인문학	3.00	2

5. 결 론

연구데이터의 재활용과 공유는 유사 연구 중복 수행 방지를 통한 연구비 절감, 과학적 연구과정 재현을 통한 연구의 투명성 제고, 데이터의 통합과 재해석을 통한 과학적 발견의 도모 등 다양한 가치를 지니고 있다. 연구데이터 공유의 중요성이 증가하면서 전 세계적으로 이와 관련된 정책과 제도가 마련되고 있으며, 데이터를 출판하기 위한 레포지토리도 증가하고 있다. 그러나 연구데이터 재활용에 있어, 무엇보다 간과할 수 없는 것은 표준적인 인용 방식을 통해 발견을 촉진하고 연구자에게 적절한 크레딧이 돌아갈 수 있도록 하는 것이다.

연구데이터 인용 데이터베이스인 DCI를 기반으로 연구데이터의 현황과 고인용 데이터의 특징을 분석한 결과를 요약하면 다음과 같다.

첫 번째, 최근 10년간 생성되어 공유되고 있는 연구데이터의 규모는 300만건을 육박하며 유전학과 생명공학이 압도적인 비중을 차지한다. 데이터를 출판하는 레포지토리로는 생명공학 분야의 Gene Expression Omnibus와 전분

야를 망라하는 Figshare가 가장 큰 규모를 보였으며 데이터의 유형으로는 데이터세트가 가장 많이 구축되어 있는 것으로 나타났다.

두 번째, 상위 인용된 500건을 대상으로 한 분석에서는 유전학보다 경제학, 사회학, 인구통계학의 비중이 더 높았으며, 전 분야를 포괄하는 UK Data Archive와 사회과학 분야인 ICPSR (Inter-University Consortium for Political and Social Research)이 가장 높은 비중을 차지하는 것으로 나타났다. 또한 고인용 연구데이터의 유형은 데이터세트보다는 데이터스터디가 많았으며, 데이터의 형식은 서베이 데이터가, 데이터 조사방법으로는 인터뷰가 가장 높은 비중을 나타냈다.

세 번째, 피인용도 횟수를 살펴본 결과, 500건 이상 인용된 데이터는 단 2건뿐인 것으로 나타났고 20-50회 사이에 238건이 분포하였다. 나머지 데이터는 상대적으로 저조한 인용 회수를 보여 인용 횟수 분포표는 긴꼬리 모양의 그래프를 나타냈다. 데이터 유형과 인용도간의 관계 분석 결과에서는 데이터스터디의 인용도가 확연히 높게 났으며, 주제와 인용도간의 분석 결과에서는 사회과학분야가 가장 높은 인용도를 보이는 것으로 나타났다. 또한 집단간 모두 통계적으로 유의한 수치를 보여, 데이터 유형과 주제가 피인용도에 영향을 주는 것으로 분석되었다.

네 번째, 고인용 연구데이터 중 DOI가 존재하는 데이터를 대상으로 알트매트릭스 분석을 수행해 본 결과, 참고문헌관리도구인 멘델리의 저장 기록만이 나타나, 연구데이터의 전체적인 사회적 영향도는 높지 않은 것으로 분석되었다. 멘델리에 저장된 연구데이터의 특성은 데이터스터디가 대다수이고 사회과학분야가 가장 많아 DCI의 분석 결과와 유사하게 나타났다.

연구데이터의 공유는 그 중요성이 전 세계적으로 인정되고 있으며, 연구의 영향력 평가 체계도 연구 부산물의 인용을 포함하는 체계로 확장이 요구되고 있다. 연구데이터 관리와 공유에 대한 인식이 부족한 우리나라에서는 먼저 개인 연구자의 연구데이터 관리를 위한 교육과 지원을 시작할 필요가 있겠으며, 연구지원기관에서도 연구에 딸린 부속물이 재활용될 수 있도록 각종 기반 마련을 서둘러야 할 것이다. 이와 더불어 데이터 출판이 가능한 레포지토리의 설치, 기관 레포지토리의 데이터 레포지토리 통합 방안 등도 포괄적으로 논의되어야 할 것이며, 이러한 기반이 마련된 후에는 연구 영향력 평가 체계의 확장에 대해서도 고민이 필요할 것이다.

참 고 문 헌

[1] 김운봉, 김용민, 양진옥. 2014.『유전체 빅데이터 연구 동향』. [online] [cited 2015. 10. 10.] 〈http://m.bioin.or.kr/board.do?num=249060&bid=report&cmd=view〉

[2] 김지현. 2014. 대학도서관의 연구데이터관리서비스에 관한 연구: 미국 연구중심대학도서관을 중심으로.『한국비블리아학회지』, 25(3): 165-189.

[3] 西薗, 由依. 2013.『オープンアクセス時代の研究成果のインパクトを再定義する: 再利用とAltmetricsの現在』. 第3回 SPARC Japan セミナー2013. [online] [cited 2015. 10. 10.] 〈http://www.nii.ac.jp/sparc/event/2013/pdf/20131025_1.pdf〉

[4] 池内, 有爲. 2014.『研究データ共有時代における図書館の新たな役割: 研究データマネジメントとデータキュレーション』. カレントアウェアネス, 319. [online] [cited 2015. 9. 10.] 〈http://current.ndl.go.jp/ca1818〉

[5] DataOne. *Education Modules* Homepage. [online] [cited 2015. 9. 10.] 〈https://www.dataone.org/education-modules〉

[6] DataCite. *DataCite Homepage*. [online] [cited 2015. 11. 15.] 〈https://www.datacite.org〉

[7] DataCite. 2015. *DataCite Metadata Schema for the Publication and Citation of Research Data*. [online] [cited 2015. 8. 15.] 〈https://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf〉

[8] Department for Business, Innovation & Skills Prime Minister's office. 2013. *G8 Science Ministers Statement London UK*. [online] [cited 2015. 8. 15.] 〈https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf〉

[9] Force, M. M. and Auld, D. M. 2014. "Data Citation Index: Promoting Attribution, Use and Discovery of Research Data." *Information Services and Use*, 34: 97-98.

[10] Force, M. M. and Robinson, N. J. 2014. "Encouraging Data Citation and Discovery with the Data Citation Index." *J Comput Aided Mol Des*, 28: 1043-1048. [online] [cited 2015. 8. 15.] 〈DOI: 10.1007/s10822-014-9768-5〉

[11] Haustein, S., Costas, R. and Larivière, V. 2015. "Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns." *PLoS ONE*, 10(3): e0120495.

[12] Havard Library. Citing Your Data Homepage. [online] [cited 2015. 8. 10.] 〈http://isites.harvard.edu/icb/icb.do?keyword=k78759&pageid=icb.page415671〉

[13] Mohammadi, E. and Thelwall, M. 2014. "Mendeley Readership Altmetrics for the Social Sciences and Humanities: Research Evaluation and Knowledge Flows." *Journal of the Association for Information Science and Technology*, 65(8): 1627-1638.

[14] National Science Foundation. 2012. *Issuance of a New NSF Proposal & Award Policies and Procedures Guide*. [online] [cited 2015. 8. 15.] 〈http://www.nsf.gov/pubs/2013/nsf13004/nsf13004.jsp〉

[15] OECD. 2007. OECD *Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD Publication. [online] [cited 2015. 9. 10.] 〈http://www.oecd.org/sti/sci-tech/38500813.pdf〉

[16] Sayogo, D. S. and Pardo, T. A. 2013. "Exploring the Determinants of Scientific Data Sharing: Understanding the Motivation to Publish Research Data." *Government Information Quarterly*, 30(1): S19-S31.

[17] Torres-Salinas, D., Martín-Martín, A. and Fuente-Gutiérrez, E. 2014. "Analysis of the Coverage of the Data Citation Index-Thomson Reuters: Disciplines, Document Types and Repositories." *Revista Española de Documentación Científica*, 37(1): 1-6. [online] [cited 2015. 9. 10.] 〈DOI: dx.DOI.org.proxy.lib.umich.edu/10.3989/redc.2014.1.1114〉

[18] The Office of Science and Technology Policy. 2013. *Increasing Access to the Results of Federally Funded Scientific Research*. Washington, D.C. [online] [cited 2015. 8. 15.] 〈https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf〉

[19] Thomson Reuters. 2015. "Data Citation Index, 2 November 2015". Personal Communication.

[20] Zahedi, Z., Costas, R. and Wouters, P. 2014. "How Well Developed Are Altmetrics? A Cross-Disciplinary Analysis of the Presence of 'Alternative Metrics' in Scientific Publications." *Scientometrics*, 101(2): 1491-1513.

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

[1] Kim, U. B., Kim, Y. M. and Yang, J. O. 2014. *Study on Trend of Research about Genom Bigdata*. [online] [cited 2015. 10. 10.] 〈http://m.bioin.or.kr/board.do?num=249060&bid=report&cmd=view〉

[2] Kim, Jihyun. 2014. "A Study on Research Data Management Services of Research University Libraries in the U.S." *Journal of Korea Biblia Society for Library and Information Science*, 25(3): 165-189.