문헌정보학 분야 연구데이터 공유에 관한 연구*

A Study on the Sharing of Research Data in Library and Information Science Field

조재인 (Jane Cho)**

초 록

본 연구는 Figshare를 통해 공유되고 있는 문헌정보학분야 연구데이터의 유형, 주제, 공개 수준 등을 분석하고 재사용성이 상대적으로 높은 데이터의 특성을 통계적으로 해석해 보았다. 분석 결과 데이터의 유형은 dataset과 paper 유형이, 주제 분야는 open access와 research data가 가장 많은 비중을 차지하였으며, 70%에 가까운 연구데이터가 pdf와 같이 편집과 재사용이 원활하지 않은 형태로 공개되어 있는 것으로 조사되었다. 또한 연구데이터의 특성과 활용 정도간의 관계 분석 결과, 주제에 있어서는 APC(Article Processing Charge)를 비롯한 open access 영역이 가장 많이 활용되고 있는 것으로 나타났으며, 데이터 유형에 있어서는 paper의 활용도가 가장 높은 것으로 나타났다.

ABSTRACT

This study analyzed the type, subject and open level of research data in the field of library and information science field shared by Figshare, and statistically analyzed the characteristics of data with relatively high recyclability. The results of the analysis showed that datasets and papers were most common data types, and open access and research data were the most common keywords of data, and that 70% of the data were published in a form that can not be processed mechanically such as pdf. As a result of analysis of the relationship between characteristics of research data and degree of sharing, open access areas such as APC (Article Processing Charge) were found to be most common in the subject. However in data type, gray literature such as paper found to be highly utilized rather than dataset.

키워드: 연구데이터, 오픈데이터, Figshare, 오픈사이언스, 문헌정보학
research data, open data, Figshare, open science, LIS

* 본 연구는 2017년도 인천대학교 자체연구비지원으로 수행되었음.
** 인천대학교 문헌정보학과 부교수(chojane123@naver.com)
■ 논문접수일자 : 2017년 11월 16일 ■ 최초심사일자 : 2017년 12월 1일 ■ 게재확정일자 : 2017년 12월 15일
■ 정보관리학회지, 34(4), 59-79, 2017. 〔http://dx.doi.org/10.3743/KOSIM.2017.34.4.059〕

1. 서 론

연구데이터의 공개는 연구의 효율성과 과정상의 투명성을 제고할 뿐 아니라, 데이터의 재사용과 통합을 통해 새로운 과학적 발견을 유도할 수 있다(Sayogo & Pardo, 2013). G8 과학 장관 회의에서 합의된 오픈데이터헌장(Department For Business, Innovation & Skills Prime Minister's Office, 2013)을 계기로 연구데이터 공유에 대한 논의가 지속되는 가운데, 각국에서도 관련 정책들이 수립되고 있다. 미국 백악관과학기술정책국은 연방 예산에 의한 연구 성과물, 즉 피어 리뷰 출판물과 연구데이터가 공공에게 오픈 액세스될 수 있도록 하는 정책의 기초를 천명하였으며(The office of Science and Technology Policy, 2013), 그에 따라 상무부, 항공우주국 등의 연방정부기관에서는 연구자들에게 데이터관리계획(Data Management Plan)의 제출을 의무화하고 있다(CENDI, 2017). 영국의 RCUK(Research Councils UK), 일본의 과학기술진흥원(Japan Science and Technology Agency) 등에서도 오픈 사이언스 기조하에 유사한 정책들이 개발되고 있어 주목되고 있다. 이러한 조류에 의해 RDA(Research Data Alliance)와 같은 국제 단체가 등장해 연구데이터 공유와 재사용을 위한 기반을 다져가고 있으며 그동안 학술 커뮤니티에서 보편적 규범으로 다루어지지 않았던 연구데이터의 표준적 인용을 위한 노력도 가속화되고 있다.

한편, 앞서 말한 연구비 조성 기관의 데이터 공개 의무화가 잇따르면서, 많은 대학 도서관에서 연구데이터 지원 서비스가 시작되고 있으며 연구데이터를 아카이빙하기 위한 데이터 레포지토리의 개발과 설치도 활발해지고 있다. 대학뿐 아니라, 출판사, 연구 지원 기관, 데이터 포털 등에서도 레포지토리를 마련하고 있으며 Figshare와 같은 클라우드 기반의 자기 공개형 데이터 레포지토리도 활성화되고 있다.

연구데이터는 그동안 주로 천문학, 지구 과학, 유전체학, 소립자 물리학 등의 분야를 통해 공유되면서 과학적 문제 해결에 눈부신 성과를 올려왔다. 그러나 최근에는 위와 같은 조류에 의해 자연 과학 분야뿐 아니라, 인문・사회 과학 분야에서도 연구데이터 공유와 재사용에 대한 움직임이 시작되고 있다. 이러한 맥락에서 본 연구는 문헌정보학 분야에서는 과연 어떠한 주제 분야와 특성을 가진 데이터가 연구자들 사이에서 공유되고 있으며 어떠한 데이터가 재사용성이 높은지를 확인해 보고자 한다. 전 주제 분야를 다루는 데이터 레포지토리 중 가장 큰 규모를 가진 Figshare를 대상으로 문헌정보학 분야의 연구데이터를 추출해 데이터 유형, 주제 분야, 공유 정도와 공개 수준을 분석해 보며, 또한 특별히 많은 연구자들에게 공유되고 있는 연구데이터가 가진 특성을 분석해 보고자 한다.

2. 이론적 배경

2.1 연구데이터 공유 정책과 데이터 레포지토리

대용량 유전체 데이터는 유전체연구의 방법론적 혁명을 통해 생명공학 기술 패러다임의 변화 뿐 아니라, 신약, 진단, 환경, 농축산 등 모든 산업에 파급 효과를 미치고 있다(김운봉, 김용

민, 양진옥, 2014). 생명 과학 분야에서 뿐 아니라, 점점 다양한 학문 분야에서 연구데이터의 공유와 재사용이 시작되고 있는데, 데이터 공유의 의의는 이렇게 연구 결과의 재현과 검증을 가능하게 함으로써 연구의 투명성을 향상시킬 뿐 아니라, 재사용과 협력을 통해 과학적 발전을 촉진시키는데 있다.

연구데이터에 대한 OECD 원칙과 가이드라인에서는 공적 자금 지원을 받은 연구데이터는 공적 관심 중에 생산된 공공재로 지적 소유권을 침해하지 않는 한 시의 적절하고 제한 없이 이용 가능해야 한다고 천명하고 있다(Organizatioon for Economic Co-operation and Development, 2007). 미국에서는 2011년 NSF와 전미인문과학기금에서 연구 자금 신청을 하는 연구자들에게 데이터 관리 계획(Data Management Plan) 제출을 요구하였고, 2013에는 백악관과학기술정책국에서 연구 개발 지출액이 연간 $100 million이 넘는 정부 기관에 대해 6개월 이내에 연구 성과의 퍼블릭 액세스를 확대하기 위한 계획안을 제출하도록 명령하였다(The office of Science and Technology Policy, 2013). 영국 RCUK(Research Councils UK)는 데이터 정책에 대한 일반 원칙(Research Councils UK, 2015)에 의해 가능한 한 제약 없이 적시에 책임 있는 방식으로 데이터를 공개하도록 요구하였다. 구체적으로 생성자의 지적 기여를 인정할 수 있는 정보를 포함하여 발견과 접근이 가능한 충분한 메타데이터가 작성되어야 하며, 원데이터 자체는 재활용이 가능한 형태로 공개되어 한다고 하였다. 일본 JST는 2017년 4월 공공 자금이 투입된 연구를 대상으로 데이터 관리계획(DMP) 수립을 권고하는 오픈 사이언스 정책을 발표하였으며 일부 기관에서 시범 운영하고 있다(Japan Science and Technology Agency, 2017). 우리나라에서도 한국과학기술정보연구원(KISTI)과 한국연구재단에서 연구의 원자료를 수집하여 공유하는 시스템을 운영하고 있다.

한편, 분야, 국가 등 다양한 장벽을 넘어 데이터를 공유하기 위해서는 영구 디지털 식별자, 메타데이터 프레임웍과 같은 기술과 제도적 기반 마련이 필요하다. 따라서 이러한 활동을 지지하는 개인 및 기관이 모여 RDA(Research Data Alliance)가 창설되었다(篠田麻美, 2014). RDA는 2017년 8월 기준 129개국 5,900개 기관이 참여하는 분야를 초월한 연구데이터 동맹으로 연구데이터 활용과 영향력 측정 활성화를 위해 기술 및 정책 기반 마련에 노력하고 있다. 또한 데이터를 재사용하고 영향력을 추적함으로써, 데이터 생산자에 대한 기여 인정 체계를 만드는 것을 목표로 DataCite(https://www.datacite.org/whycitedata)가 창설되었으며 Clarivate analytics는 Data Citation Index를 통해 수백 개의 공신력 있는 데이터 레포지토리의 메타데이터를 인덱스하고 인용 정보를 제공하고 있다.

한편, 대표적인 데이터 레포지토리 등록소인 re3data.org를 살펴보면, 현재 전 세계 67개 국가에 1,900개 이상의 레포지토리가 등록되어 있음을 확인할 수 있다. 1,900여개의 레포지토리 중 1,600여 개는 오픈엑세스 형식으로 서비스되고 있으며 80여 개 정도는 데이터의 발견과 접근, 재사용성 등에 있어 DSA(Data Seal of Approval), WDS(The ICSU World Data System)등의 인증을 받고 있다. 레포지토리 소프트웨어는 dSpace와 같은 기관 레포지터리 시스템이나 하버드대학이 배포하고 있는 Dataverse가 가장

많이 사용되고 있으나 최근 캘리포니아 공과대학이 개발한 Caltech DATA와 같은 전용 레포지토리나, 캐나다 연구도서관 협의회(CARL)의 Federated Research Data Repository(FRDR), 일본 NII(National Institute of Informatics)가 계획하고 있는 클라우드 기반의 데이터 레포지토리처럼 대학도서관이 공동으로 활용하는 연합형 플랫폼의 개발도 주목되고 있다(科學技術・學術審議會 學術分科會 學術情報委員會, 2015). 그 밖에 데이터 저널, 프로젝트 기관의 웹 사이트를 통해 연구데이터가 출판되기도 하며, arXiv, GenBank, ICPSR, Dryad와 같은 주제별 저장소, Figshare 같은 클라우드 기반의 아카이브를 통해 자기 공개되기도 한다.

Figshare는 연구자가 실험에서 실패하거나 공개되지 않은 데이터 등을 등록, 공개하고 다른 연구자들과 공유함으로써 타 연구자가 비슷한 실험을 하는데 드는 시간과 경비를 절감하기 위하여 시작되었다. 연구데이터를 9개의 데이터 유형(F: figure, M: media, D: dataset, F: fileset, Po: poster, Pa: paper, Pr: presentation, T: thesis, C: code)으로 구분해 연구과정에서 산출된 다양한 유형의 부산물을 등록할 수 있도록 하고 있다. 전통적인 학술출판방식이 아닌 새로운 방식으로 연구 성과를 배포하기 위한 박사과정 학생의 실험으로 출발하였으며 현재는 Digital Science사에 의해 운영되고 있다. Data Citation Index에 등록된 다학제적 분야 레포지토리 중 가장 방대한 데이터량을 보이고 있는 Figshare는 COPE(The Committee on Publication Ethics)의 Principles of Transparency and Best Practice in Scholarly Publishing 정책을 따르고 있으며 라이선스는 CC(Creative Commons)를 채택하고 있다. 또한 DataCite Metadata Schema, OAI PMH를 통해 호환성을 갖추고 있으며 출판과 동시에 DataCite DOI를 받을 수 있어, 전통적 인용 방식처럼 연구데이터가 인용될 수 있다. 한편 Figshare는 PLOS, Springer-Nature와 제휴해 연구 결과에 대한 재현성을 검증하는 프로젝트(Reproducibility Initiative)에도 참여하고 있어 주목된다. 이미 나온 연구 결과에 대하여 추가 실험을 실시하여 과학자의 연구 결과에 대한 재현성을 인정하는 이 프로젝트에서 Figshare는 논문의 부속자료를 호스팅하는 역할을 맡고 있다(Hahnel, 2012).

2.2 선행연구

국내에서 연구데이터에 관한 연구는 연구데이터의 관리 및 공유에 대한 연구자들의 인식(강주연, 2017; 김은정, 남태우, 2012; 김지현, 2012)이 가장 활발하게 다루어지고 있으며, 해외 사례조사 및 향후 국내 실행방향을 제시하는 연구(김지현, 2013; 심원식, 2016)도 지속적으로 이루어지고 있다. 또한 최근에는 연구데이터의 인용(김지현, 정은경, 윤정원, 이재윤, 2017; 조재인, 2016)에 관한 연구도 수행되고 있으며, KRM(Korea Research Memory)을 중심으로 인문사회 아카이브의 발전 방향(신영란, 정연경, 2012; 심원식, 안혜연, 변제연, 2015)에 대해서도 논의되고 있다.

최근에는 DCI(Data Citation Index)를 통해 연구데이터의 인용 정도를 측정하거나 Altmetrics를 통해 연구데이터의 사회적 영향력을 파악하는 연구도 이루어지고 있다. DCI를 통해 데이터의 인용도를 분야별로 분석한 연구들을 종합해 보면

(조재인, 2016; Torres-Salinas, Martín-Martín, Fuente-Gutiérrez, 2014), 연구데이터의 85%는 인용되지 않았지만 인용이 증가하는 추세이며, 유전학과 생명공학 분야가 가장 큰 규모를 보이지만, 인구, 고용 등 경제・사회과학분야의 데이터가 상대적으로 많이 인용되고 있음을 파악할 수 있다. 한편, 아직까지 고유 식별자가 부여되지 않아 Altmetrics 측정이 불가한 연구데이터가 다수 존재하지만(Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2016), 고인용 연구데이터 중 일부를 추출하여 Altmetrics 지수를 비교한 결과, 사회과학 분야 데이터 스터디에 대한 지수가 상대적으로 높게 나타났다는 연구 결과도 제시된 바 있다(조재인, 2016).

특정 분야를 기반으로 해당 분야 연구데이터의 특성이나 주제를 분석한 연구는 많지 않지만, 다음과 같이 몇 가지 연구 결과를 제시해 볼 수 있겠다. 사회학 분야를 대상으로 연구데이터의 지적 구조를 분석한 최형욱과 정은경(2017)은 DCI에 구축된 사회학 분야 연구데이터의 주제어를 기반으로 동시출현단어 분석과 저자 네트워크 분석을 수행하였다. 그 결과 전통적인 논문 기반의 지적 구조와 달리 공중보건과 심리학의 중심성이 높음을 밝히고 있다. 또한 문헌정보학분야로 한정해 연구데이터 공유에 관해 분석한 연구도 다음과 같이 찾아 볼 수 있다. Borrego와 Garcia(2013)는 ISI Journal Citation Reports에서 문헌정보학 저널을 추출해 저장하고 있는 연구데이터의 특성을 분석하였다. 이 연구는 데이터 레포지토리를 대상으로 하지 않고 온라인 저널을 대상으로 하였다는 점에서 본 연구와 차이가 있으나 문헌정보학 분야에서 생산된 연구데이터의 특성을 분석하였다는 점에서 유사성이 있다고 말할 수 있다. 그들은 논문에 부속된 데이터는 주로 방법론에 대한 추가적인 설명이거나 발표된 논문에는 제시되지 않은 분석 결과라고 밝혔으며, 일반적으로 pdf나 워드 파일 형태를 가지고 있다고 설명하였다. 또한 Aleixandre-Benavent, Moreno-Solano, Sapena, Sánchez Pérez(2016)는 Web of Science database에서 문헌정보학 분야의 저널을 대상으로 연구데이터 공개 정책과 영향력 지수와의 관계를 분석하였는데, 문헌정보학분야 60% 이상의 저널이 연구데이터의 재활용을 허용하고 있으며, JCR 순위가 높은 저널일수록 오픈 정책이 활성화되어 있다고 밝혔다.

한편, 본 연구에서 분석 대상으로 하고 있는 Figshare를 대상으로 연구데이터의 특성이나 공유 정도를 분석한 선행연구는 많지 않지만 Figshare에 등재된 연구데이터의 공유 카운트를 통해 전반적인 재활용성을 파악한 연구가 수행된 바 있다(Thelwall & Kousha, 2016).

3. 연구의 방법

문헌정보학 분야의 주제 저장소는 존재하지 않기 때문에 다학제적 레포지토리 중 가장 구축된 데이터량이 방대한 Figshare를 대상으로 연구데이터를 추출하여 분석을 수행하였다. Figshare는 연구데이터를 9개의 데이터 유형(F: figure, M: media, D: dataset, F: fileset, Po: poster, Pa: paper, Pr: presentation, T: thesis, C: code)으로 구분하고 있다. poster, code, media 등 연구 과정에서 산출된 다양한 유형의 부산물과 paper, these 형태의 회색문헌도 연구데이터

의 범위에 포함시키고 있다. 연구데이터의 정의와 범위는 기관마다 조금씩 다르지만, DataCite (2012)에서는 관찰, 실험, 경험에 기반하는 사실 데이터뿐 아니라 통계기록, 음원, 미디어, 이미지 등 다양한 매체에 기록된 연구의 부산물로 정의되고 있다. 또한 데이터를 기술한 데이터페이퍼와 회색문헌까지도 연구데이터의 범주로 인정하여 고유 식별자가 부여되고 있어, 본 연구에서도 Figshare가 포괄하는 모든 유형의 데이터를 누락 없이 분석 대상으로 하였다.

데이터 추출 및 분석 방법을 자세히 기술하면 다음과 같다.

첫 번째, Figshare에서 주제 분야를 "category: Library and Information"으로 설정하여 데이터를 검색한 후, most shared 순으로 소팅하여 유형별로 100건씩 데이터를 추출하였다. 9개의 데이터 유형 중 100건 이상이 누적된 데이터 유형 6개에서 600건의 연구데이터를 추출하며, 100건 미만의 데이터를 포함하고 있는 영역인 media, thesis, code로부터는 구축된 데이터를 모두 추출해 총 659건을 분석 대상으로 하였다.

두 번째, 추출된 문헌정보학 분야 연구데이터를 대상으로 데이터 유형, 주제 등의 구축 현황 및 활용 정도를 분석하였다. 또한 'Five Stars of Linked Data'(http://5stardata.info/ko/) 모델의 5단계 평가 기준을 적용하여 구축된 데이터가 어떠한 단계의 공개 수준을 보이고 있는지 확인하였다. 'Five Stars of Linked Data'는 팀 버나드리가 제시한 오픈 데이터 발달 단계에 따른 측정 지표로 데이터의 개방성과 기계판독성 등을 평가기준으로 하고 있다.

세 번째, 연구데이터의 유형에 따라 어떠한 파일 형식을 채택하고 있으며, 또한 연구의 주제에 따라 어떠한 유형의 데이터가 구축되어 있는지 교차분석과 대응일치 분석을 통해 파악하였다. 대응일치분석(Correspondence Analysis)은 다차원척도법의 일종(김상수, 2005)으로 행 범주와 열범주에 대한 최적의 수량화 값을 계산하여, 자료를 2차원 평면상에 플롯팅함으로써, 변수간의 관련성을 파악하는 기법이다. 대응일치분석에서는 상호 연관성을 파악하기 위해 주로 2차원 좌표에서 점들의 분포를 보고 판단한다(류귀열, 2011).

네 번째, 연구데이터 특징과 활용 정도 간의 관계를 비모수기법인 크루스칼 왈리스(Kruskal-Wallis H)와 스피어만(Spearman) 분석을 통해 파악함으로써 어떠한 주제와 유형의 데이터가 문헌정보학 연구자에게 다수 활용되고 있으며, 연구데이터의 공개 수준과 활용 정도 간에도 어떠한 관계가 존재하는지 통계적으로 검증하였다. 그렇게 함으로써 재사용 가능성이 높은 문헌정보학 연구데이터의 특성을 이해하였다.

4. 분석 결과

4.1 연구데이터의 공유 정도 및 공개 수준 분석

Figshare의 문헌정보학 관련 연구데이터는 2016년 중반을 기준으로 총 1,583건이 구축되어 있는 것으로 확인되었다.

첫 번째, 구축된 데이터의 유형은 〈표 1〉과 같이 총 9가지로 구분되는데, dataset과 paper가 각각 360(22.74%)건으로 가장 많은 비중을 보이는 것으로 나타났다. dataset은 csv와 같은

〈표 1〉 문헌정보학 연구데이터 유형 분포

유형	개수	비중%
dataset	360	22.74
paper	360	22.74
presentation	298	18.83
figure	219	13.83
fileset	144	9.1
poster	143	9.03
thesis	34	2.15
media	13	0.82
code	12	0.76
계	1,583	100.00

원자료 형태나 xls와 같은 응용 소프트웨어 형식으로 구축되어 있는 원데이터셋을 의미하며 paper는 원데이터가 아니라 이를 분석하거나 해석한 결과물, 또는 논문 자체나 부속 자료를 비롯한 각종 회색문헌을 포함한다. 앞에서 언급한 바와 같이 DataCite(2012)에서는 연구데이터의 범주에 데이터셋, 데이터페이퍼뿐 아니라, 상업적인 출판 유통 채널을 통해 획득하기 어려운 회색문헌을 포함하고 있는데, Figshare에서도 paper 뿐 아니라 다양한 형태를 가진 회색문헌이 존재하는 것으로 분석되었다. 한편, dataset과 paper 이외에도 데이터 유형으로 presentation(298, 28.83%), figure(219, 13.83%)가 높은 비중을 차지하는 것으로 나타났으며, 비중은 낮지만 code(12건 0.76%), media(13건, 0.82%) 등의 데이터 유형이 존재하는 것으로 나타났다.

두 번째, 구축된 문헌정보학 연구데이터의 주제 분야를 살펴보기 위하여 연구 대상이 된 데이터에 부여된 키워드 총 1,944개를 추출하여 주제 분석을 수행하였다. 두 번 이상 출현된 키워드는 321개, 10번 이상 빈번하게 등장한 키워드는 총 30개로 나타났으며, 가장 많은 출현빈도를 보인 키워드는 〈표 2〉와 같이 open access(102번)로 나타났다. 그 다음 많은 출현빈도를 보인 키워드는 research data(87회)와 altmetrics(83번)로 나타나 open access나 open science 그리고 계량정보학 연구에 활용된 연구데이터가 다수 Figshare에 공개되어 있는 것으로 분석되었다. 출현빈도가 높은 키워드를 중심으로 공개된 데이터의 내용을 살펴보면 open access의 경우 APC(article processing charges) 데이터(Kiley, 2014)나 오픈엑세스 저널리스트(Crawford, 2015) 등이, altmetrics의 경우 위키피디어에 참조된 저널리스트(Halfaker & Taraborelli, 2015) 등이 존재하였다.

세 번째, 연구데이터의 공유 정도를 살펴보기 위하여 분석 대상이 된 659건의 평균 뷰(view)와 다운로드(download) 횟수를 파악한 결과를 제시하면 〈표 3〉과 같다. 뷰의 최소값은 10회, 최대값은 17,993회로 나타났으며, 평균 수치는 541회인 것으로 나타났다. 그러나 다운로드 횟수는 뷰 횟수와는 달리 최소값은 0, 최대값은 3,004회, 평균 수치는 78회를 보여, 뷰

〈표 2〉 문헌정보학 연구데이터에 부여된 고출현 키워드

키워드	출현 순위	빈도
open access	1	102
research data	2	87
altmetric	3	83
scholarly communication	4	52
article processing charges	5	47
bibliometrics	6	37
libraries	7	33
jisc	8	27
open science	9	26
citation	10	26

〈표 3〉 문헌정보학 연구데이터의 뷰, 다운로드, 트윗 평균 및 상관성 분석 결과

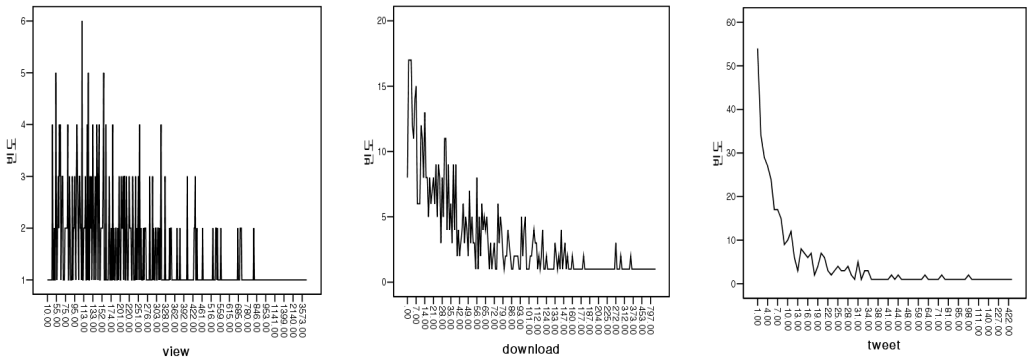
구분	N	최소값	최대값	평균
뷰	659	10.0	17,993.0	541.8
다운로드	659	.0	3,004.0	78.4
트윗	389	1.0	516.0	20.6

구분		view	download	tweet
뷰	Pearson 상관	1	.731**	.860**
	유의확률(양측)		.000	.000
다운로드	Pearson 상관	.731**	1	.676**
	유의확률(양측)	.000		.000
트윗	Pearson 상관	.860**	.676**	1
	유의확률(양측)	.000	.000	

한 경우에 비해 약 1/7배 정도 수준에 머무르는 것으로 나타났다. 한편, 소셜미디어의 노출정도를 확인하기 위하여 해당 연구데이터가 트윗에 노출된 빈도를 살펴보니, 분석 대상 데이터의 약 절반이 넘는 연구데이터가 트윗에 노출된 바 있는 것으로 나타났다. 한편, 뷰와 다운로드, 뷰와 트윗횟수, 다운로드와 트윗 횟수간의 상관성은 각각 r= 0.731, r= 0.860, r= 0.676으로 나타나 모두 상호간 강한 상관성을 나타내고 있는 것으로 분석되었다. 17,993회로 가장 다수에 의해 뷰되었을 뿐 아니라, 가장 많이 다운로드(3,004회)되고 트윗(516회)된 연구데이터가 존재하였는데, 이는 Kramer와 Bosman(2015)이 FORCE 2015에 제출한 학술커뮤니케이션 혁신과 관련된 poster 데이터로 나타났다. 두 번째로 높은 뷰횟수(15,951건)를 보인 데이터는 csv dataset으로 연구기관 리스트로 구성된 GRID release 2015(Digital-science, 2015)였으며, 역시 두 번째로 많이 다운로드(1,016회) 및 트윗(422회)된 데이터는 영국 고등교육기관이 10개의 메이져 출판사에 지급한 학술지 구독 비용이 공개되어 있는 csv 데이터(Lawson, Meghreblian, & Brook, 2015)인 것으로 나타났다. 한편, 데이터 활용 경향을 확인하기 위하여 뷰, 다운로

드, 트윗 횟수 통계를 활용하여 그래프를 그린 결과, 세 개의 그래프 모두 다수 공유된 연구데이터는 소수에 이를 뿐이고 대부분의 데이터는 활용도가 낮은 경향을 나타냈다. 이러한 경향은 〈그림 1〉과 같이 뷰에서 트윗으로 갈수록 극단적으로 발전해 긴꼬리 모양을 나타내고 있었다.

네 번째, 연구데이터의 공개 수준을 살펴보기 위하여 분석 대상이 된 659건의 데이터 형식을 'Five Stars of Linked Data' 모델의 5단계로 구분하여 통계치를 산출한 결과는 다음 〈표 4〉와 같다. 전 세계 공공데이터 개방 지수 ODB(Open Data Barometer) 등에서 참조하고 있는 'Five Stars of Linked Data' 모델은 데이터 존재여부, 개방성, 최신성, 기계판독성, 용이성 등을 평가기준으로 오픈 데이터의 공개 수준에 대한 기준을 제시하고 있다. 이 모델에서는 데이터 개방시 기계판독이 가능한 원천 데이터 개방을 가장 중요한 요건으로 꼽고 있다(백인수, 2013). 이 모델을 참고하여 연구데이터의 공유 단계를 5가지 형태로 기술하면 다음과 같다. 1단계(★OL: On-Line)는 포맷에 상관없이 데이터를 개방형 라이선스하에 공개한 상태를 의미한다. 데이터가 수록된 문건을 스캐닝해 pdf와 같은 형태로 오픈 라이선스하에 공개하고 있다면 1단계에 해당된다고 볼 수 있겠다. 2단계(★★OL, RE: machine REadable)는 특정 소프트웨어를 사용하여 처리할 수 있는 구조화된 형태를 의미한다. 가령 표에 들어있는 연구데이터를 엑셀파일과 같은 포맷으로 작성하여 공유한다면 2단계에 해당된다고 말할 수



뷰회수 다운로드 횟수 트윗 횟수

〈그림 1〉 뷰, 다운로드, 트윗 횟수 그래프

〈표 4〉 오픈데이터 공개 수준에 따른 구축 건수

구분	건수	비중(%)	형식
1단계	451	68.4	PDF, jpeg, png, mp3-4 등
2단계	90	13.7	ppt xls doc등
3단계	29	4.5	csv, txt, tsv 등
4단계	0	0	-
5단계	0	0	-

있겠다. 이는 pdf 형태로 공개된 1단계에 비하여 사용자가 데이터를 처리하거나 다른 포맷으로 가공할 수 있다는 점에서 공개 수준이 높다고 말할 수 있다. 3단계(★★★OL, RE, OF: Open Format)는 비독점적 포맷, 개방형 포맷으로 데이터를 공개하고 있는 상태를 말한다. csv나 txt와 같이 특정 소프트웨어에 종속되지 않는 비독점적 오픈 포맷으로 제공되는 방식을 의미한다. 이렇게 공개된 데이터는 이용자가 원하는 방법으로 데이터를 가공하거나 처리할 수 있다. 4단계(★★★★OL, RE, OF, URI)는 URI가 부여되어 있어 데이터 의미를 기계적으로 해석할 수 있는 단계를 의미한다. 데이터 항목의 재사용성을 제고할 수 있도록 RDF와 같은 구조화가 이루어지는 단계가 여기에 해당된다. 마지막 5단계(★★★★★OL, RE, OF, URI, LD: Linked Data)는 기능적 상호운용이 가능하도록 항목별 데이터를 제어하여 데이터의 문맥과 배경을 제공하는 형태이다. 링크드 오픈 데이터 방식으로 데이터 레지스트리를 통해 의미와 연관된 지식까지도 획득할 수 있는 단계라고 말할 수 있겠다.

판별하기 어려운 압축 파일 등을 제외하고 본 연구의 분석 대상 데이터를 5단계로 구분해 보면 〈표 4〉와 같이 70%에 가까운 데이터가 가공할 수 없는 상태로 공개되어 있는 것으로 나타났다. 단지 오픈라이선스 방식을 취할 뿐, 많은 양의 데이터가 편집과 재사용이 원활하지 않은 이미지 형식으로 공개되어 있어 1단계 수준에 머물러 있다는 것이다. 한편 엑셀과 같이 구조화된 2단계 수준의 공개 데이터는 90건으로 13.7%를 차지하고 있었으며, 비독점적 오픈 포맷인 csv 형식 등으로 데이터를 공개한 3단계의 데이터는 단 29건으로 4.5%에 불과한 것으로 나타났다. 한편, 개체의 의미 해석을 위해 URL를 부여하거나 LOD 형태로 구조한 4, 5단계의 데이터는 발견하지 못하였다. 종합해 볼 때, 아직까지 Figshare 공개된 문헌정보학 분야 연구데이터는 재활용성보다는 단지 오픈라이선스로 데이터 자체를 공개하는데 의미를 둔 경우가 다수를 차지하는 것으로 설명할 수 있겠다.

4.2. 연구데이터 유형과 형태, 주제 간의 관계 분석

여기에서는 분석 대상으로 추출된 문헌정보학 분야 연구데이터를 대상으로 유형과 파일형태, 그리고 유형과 주제 간에 주목할 만한 관계가 존재하는 지 여부를 교차 분석 및 대응일치 분석을 통해 조사한 결과를 제시해 보겠다.

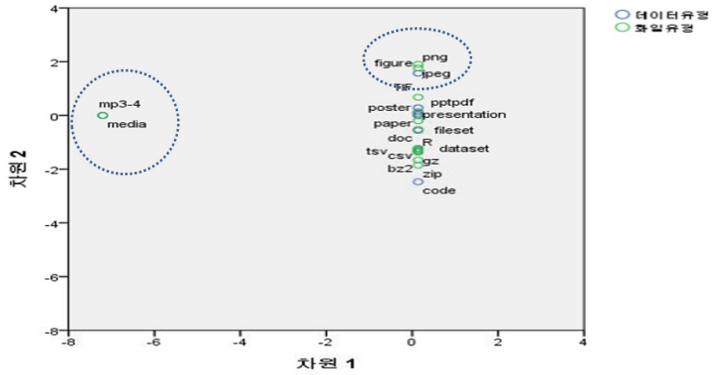
첫 번째, 데이터 유형과 파일 형식간의 관계를 교차분석을 통해 분석한 결과 연구데이터의 유형에 따라 파일 형식이 다르게 나타나는 것(χ^2 = 1605.434, p = .000)으로 검증되었다. 〈표 5〉의 교차표를 살펴보면 웨비나(Webinar), 인터뷰 등을 다루고 있는 media의 경우 대부분 mp3-4형태로 구축되어 있었으며, 자바 클라이언트, 엑셀 매크로 등의 code는 대부분 압축파일 형태를 가지고 있는 것으로 나타났다. 학술출판의 재정 플로우 모델, 연구데이터 공유 모델 등 다양한 주제 분야의 개념 설명을 위한 figure는 jpg(31.6%)나 png(30.6%)와 같은 이미지 파일로 구축된 경우가 많았으나 이러한 파일들이 pdf(33.7%) 형식으로 변환되어 공유되는 경우도 다수 존재하였다. 한편, APC 데이터, 공공 오픈데이터 리스트 등이 구축된 dataset은

〈표 5〉 데이터 유형과 파일 형식간의 교차표

구분	bz2	csv	doc	gz	jpeg	mp3-4	png	ppt	R	TIF	tsv	txt	xls	zip	pdf	계
code	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	8
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	100.0%
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	29.6%	0.0%	1.3%
data set	3	11	5	2	0	0	1	0	2	0	2	3	28	13	25	95
	3.2%	11.6%	5.3%	2.1%	0.0%	0.0%	1.1%	0.0%	2.1%	0.0%	2.1%	3.2%	29.5%	13.7%	26.3%	100.0%
	100.0%	52.4%	17.2%	66.7%	0.0%	0.0%	2.9%	0.0%	66.7%	0.0%	66.7%	60.0%	70.0%	48.1%	6.8%	15.0%
figure	0	0	1	0	31	0	30	0	0	2	0	0	1	0	33	98
	0.0%	0.0%	1.0%	0.0%	31.6%	0.0%	30.6%	0.0%	0.0%	2.0%	0.0%	0.0%	1.0%	0.0%	33.7%	100.0%
	0.0%	0.0%	3.4%	0.0%	88.6%	0.0%	88.2%	0.0%	0.0%	50.0%	0.0%	0.0%	2.5%	0.0%	9.0%	15.4%
fileset	0	10	13	1	0	0	2	13	1	2	1	2	11	3	34	93
	0.0%	10.8%	14.0%	1.1%	0.0%	0.0%	2.2%	14.0%	1.1%	2.2%	1.1%	2.2%	11.8%	3.2%	36.6%	100.0%
	0.0%	47.6%	44.8%	33.3%	0.0%	0.0%	5.9%	26.0%	33.3%	50.0%	33.3%	40.0%	27.5%	11.1%	9.3%	14.6%
media	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	12
	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%
paper	0	0	8	0	0	0	0	0	0	0	0	0	0	2	87	97
	0.0%	0.0%	8.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.1%	89.7%	100.0%
	0.0%	0.0%	27.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	7.4%	23.8%	15.3%
poster	0	0	0	0	4	0	1	1	0	0	0	0	0	0	94	100
	0.0%	0.0%	0.0%	0.0%	4.0%	0.0%	1.0%	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	94.0%	100.0%
	0.0%	0.0%	0.0%	0.0%	11.4%	0.0%	2.9%	2.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.7%	15.7%
presen-tation	0	0	0	0	0	0	0	36	0	0	0	0	0	1	62	99
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	36.4%	0.0%	0.0%	0.0%	0.0%	0.0%	1.0%	62.6%	100.0%
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	72.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.7%	16.9%	15.6%
thesis	0	0	2	0	0	0	0	0	0	0	0	0	0	0	31	33
	0.0%	0.0%	6.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	93.9%	100.0%
	0.0%	0.0%	6.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8.5%	5.2%
계	3	21	29	3	35	12	34	50	3	4	3	5	40	27	366	635
	0.5%	3.3%	4.6%	0.5%	5.5%	1.9%	5.4%	7.9%	0.5%	0.6%	0.5%	0.8%	6.3%	4.3%	57.6%	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

χ^2 = 1605.434, p = .000

주로 xls(29.5%)이나 원 데이터 형식인 csv (11.6%)로 구축되어 있으나, 이 경우에도 pdf (25%)로 변환해 공개함으로써 재활용성을 저해하는 경우가 다수 존재하였다. 그 밖에 fileset, poster, paper, presentation의 경우도 많은 데이터가 pdf로 변환되어 공개되고 있는 것으로 나타났다. 컴퓨터 환경에 관계없이 같은 표현을 하기 위한 목적으로 개발된 pdf는 장치 독립성 및 해상도 독립성을 가질 뿐 아니라, 암호화 및 압축 기술을 통해 변조가 어려우며 사용권을 다양하게 부여할 수 있다. 그러나 pdf는 이용자가 기계적으로 데이터를 추출하여 편집・가공하기 원활하지 않으므로 기계판독 가능성이 가장 중요한 요건인 오픈 데이터 형식으로는 적절하지 않다. 그럼에도 불구하고 Figshare에 구축된 문헌정보학 연구데이터는 code와 media 유형을 제외하면, 최종적으로 pdf 형식으로 변환하여 공개된 경우가 많은 것으로 나타났다. 한편, 대응분석을 통해 데이터 유형과 파일 형식간의 관계를 시각화한 결과, 〈그림 2〉와 같이 1차원



〈그림 2〉 데이터 유형과 파일 형식간 대응분석 결과

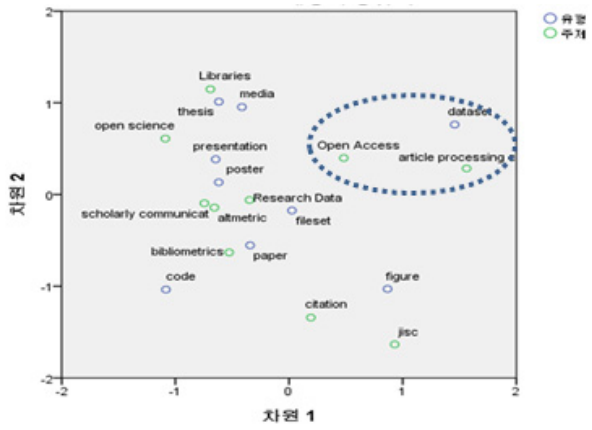
에서 39%, 2차원에서 22%로 총 61%의 누적 설명력을 보이는 지각도가 나타났다. 대응분석에서는 일반적으로 70% 이상이 되어야 설명력이 높다고 말할 수 있기 때문에(채희원, 2017) 플로팅 결과가 높은 설명력을 보인다고 말하기는 어려우나 위에서 제시한 교차분석의 데이터를 지각지도를 통해 직관적으로 파악할 수 있다. 〈그림 2〉의 플로팅 결과를 살펴보면 좌측에 media와 mp3-4, 우측 상단에 figure와 png, jpeg 등은 상호 근접하게 위치하고 있음을 확인할 수 있다. 그러나 dateset, fileset, paper, presentation, code 등의 데이터는 pdf, csv, doc와 같은 다양한 종류의 화일과 관련성을 맺고 있어, 데이터 유형과 파일 형식간의 관계에 대한 식별이 시각적으로 명확하지 않다. 특히 fileset, poster, paper 뿐 아니라 재활용이 가능한 dataset 조차 pdf 형식으로 변환하여 공개하고 있는 경우가 다수 존재해 데이터 유형과 형식간의 관계 식별을 더욱 모호하게 하고 있다.

두 번째, 연구데이터 유형과 주제 간의 관계를 교차분석을 통해 살펴본 결과에서도 통계적으로 의미 있는 차이가 존재하는 것(χ^2 = 254.423, p = .000)으로 검증되었다. 교차 분석을 위해 필요한 데이터는 앞의 〈표 1〉에서 제시한 고출현빈도 키워드 10개가 부여된 데이터 343건에서 추출하여 분석을 수행하였음을 밝힌다. 분석 결과, 가장 주의 깊게 살펴볼 필요가 있는 원데이터 형식인 dataset은 open access영역(48.9%)에 가장 많이 구축되어 있는 것으로 나타났으며, 그 다음은 38.3%로 article processing charges 영역에 다수의 dataset이 구축되어 있는 것으로 나타났다. 반면 research data, altmetric, bibliometrics, libraries 등의 영역에 분포된 데이터에는 소수의 dataset 만이 구축되어 있는 것으로 나타났다. research data 영역은 fileset(32.2%) 유형이, altmetric(38.6%)와 bibliometrics(34.6%)는 paper 유형이, libraries 영역은 thesis(33.3%)가 가장 많은 비중을 차지하고 있는 것으로 나타나 dataset 유형과 주목할 만한 관련성을 형성하고 있는 영역은 광범위하지 않은 것으로 요약할 수 있겠다. 마찬가지로 〈그림 3〉과 같이 데이터 유형과 주제 분야 간의 관계를 시각적으로 확인하기 위하여 대응분석을 수행한 결과, 1차원에서 40.2%, 2차원에서 18.2%, 총 59%

〈표 6〉 데이터 유형과 주제 분야간의 교차표

구분	Open Access	Research Data	altmetric	scholarly communication	article processing charges	biblio metrics	Libraries	jisc	open science	citation	계
code	0	0	1	0	0	1	0	0	0	0	2
	0.0%	0.0%	50.0%	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	0.0%	100.0%
	0.0%	0.0%	1.8%	0.0%	0.0%	3.8%	0.0%	0.0%	0.0%	0.0%	.6%
data set	23	2	2	0	18	0	1	1	0	0	47
	48.9%	4.3%	4.3%	0.0%	38.3%	0.0%	2.1%	2.1%	0.0%	0.0%	100.0%
	27.4%	3.4%	3.5%	0.0%	51.4%	0.0%	4.2%	7.1%	0.0%	0.0%	13.7%
figure	9	5	1	0	6	1	2	8	0	6	38
	23.7%	13.2%	2.6%	0.0%	15.8%	2.6%	5.3%	21.1%	0.0%	15.8%	100.0%
	10.7%	8.5%	1.8%	0.0%	17.1%	3.8%	8.3%	57.1%	0.0%	37.5%	11.1%
fileset	16	19	6	4	6	8	1	2	0	2	64
	25.0%	29.7%	9.4%	6.3%	9.4%	12.5%	1.6%	3.1%	0.0%	3.1%	100.0%
	19.0%	32.2%	10.5%	18.2%	17.1%	30.8%	4.2%	14.3%	0.0%	12.5%	18.7%
media	1	2	0	0	0	0	1	0	0	0	4
	25.0%	50.0%	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	100.0%
	1.2%	3.4%	0.0%	0.0%	0.0%	0.0%	4.2%	0.0%	0.0%	0.0%	1.2%
paper	10	5	22	8	5	9	0	2	1	5	67
	14.9%	7.5%	32.8%	11.9%	7.5%	13.4%	0.0%	3.0%	1.5%	7.5%	100.0%
	11.9%	8.5%	38.6%	36.4%	14.3%	34.6%	0.0%	14.3%	16.7%	31.3%	19.5%
poster	7	14	8	3	0	4	6	0	0	2	44
	15.9%	31.8%	18.2%	6.8%	0.0%	9.1%	13.6%	0.0%	0.0%	4.5%	100.0%
	8.3%	23.7%	14.0%	13.6%	0.0%	15.4%	25.0%	0.0%	0.0%	12.5%	12.8%
presen-tation	12	12	12	5	0	2	5	1	5	0	54
	22.2%	22.2%	22.2%	9.3%	0.0%	3.7%	9.3%	1.9%	9.3%	0.0%	100.0%
	14.3%	20.3%	21.1%	22.7%	0.0%	7.7%	20.8%	7.1%	83.3%	0.0%	15.7%
thesis	6	0	5	2	0	1	8	0	0	1	23
	26.1%	0.0%	21.7%	8.7%	0.0%	4.3%	34.8%	0.0%	0.0%	4.3%	100.0%
	7.1%	0.0%	8.8%	9.1%	0.0%	3.8%	33.3%	0.0%	0.0%	6.3%	6.7%
계	84	59	57	22	35	26	24	14	6	16	343
	24.5%	17.2%	16.6%	6.4%	10.2%	7.6%	7.0%	4.1%	1.7%	4.7%	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

χ^2 = 254.423, p = .000



〈그림 3〉 데이터 유형과 주제 분야간 대응분석 결과

의 누적 설명력을 보이는 지각도가 생성되었다. dataset 유형과 open access, article processing charge가 우측 상단 1사분면에 상호 근접한 위치에 자리 잡고 있으며, 그 밖의 주제 분야와 데이터 유형은 지도상에 분산되어 있는 것을 확인할 수 있다. 2차원 플로팅 결과 역시 dataset 형태를 가지고 공개되어 있는 주제 영역의 데이터가 한정되어 있음을 보여주고 있다.

4.3 데이터의 특성과 활용 정도간의 관계 분석

앞에서는 문헌정보학 연구데이터의 유형과 파일형식 그리고 주제 간에 존재하는 관련성에 대하여 살펴보았다. 여기에서는 분석 대상으로 추출된 연구데이터의 유형, 주제, 공개 수준이 활용 정도와 어떠한 관계가 있는지 통계적으로 살펴봄으로써, 재활용 가능성이 높은 연구데이터의 특성을 파악해 보도록 한다.

첫 번째, 연구데이터의 유형과 활용 정도 간의 관계를 파악하기 위하여 기술분석과 비모수 기법인 크루스칼 왈리스 분석을 수행해 본 결과, 〈표 7〉과 같이 데이터의 유형에 따라 뷰 및 다운로드 횟수에는 유의한 차이(p=.000)가 존재하는 것으로 분석되었다. 데이터를 뷰한 횟수에서는 dataset이 평균 1,092건으로 가장 높았고 그 다음 paper(804건)로 나타났으나, 연구데이터를 다운로드한 회수를 살펴보면 paper(160)가 dataset(148)보다 높은 수치를 보이는 것으로 나타났다. 이렇게 다운로드한 횟수에서 dataset보다 paper의 수치가 높은 것은 본격적인 활용의도를 가진 이용자들이 아직까지는 원자료인 dataset보다 데이터가 해석된 paper에 의존하는 경우가 더 많기 때문인 것으로 추정해 볼 수 있겠다. Figshare에 공개되어 있는 dataset은 종류가 제한적이고 변수 및 조사방법론 등에 대한 자세한 설명이 부족하거나 표준화되어 있지 않은 경우가 많다. 그러나 발표된 논문의 요약자료나 그 자체, 또는 부속자료를 제공하고 있는 paper는 데이터에 대한 자세한 설명이 포함되어 있어 원자료를 획득해 재활용하기보다 연구 결과 자체를 참고하거나 인용하기 위한 목적의 이용자에 의해 더욱 빈번하게 활용될 수 있기 때문인 것으로 보인다.

두 번째, 주제 분야에 따라 연구데이터의 활용 정도가 다르게 나타나는지 파악하기 위하여 뷰와 다운로드 횟수를 기반으로 크루스칼 왈리스 분석을 수행한 결과, 〈표 8〉과 같이 유의한 차이(p=.001, p=.000)가 존재하는 것으로 분석되었다. 뷰의 경우에는 article processing charges(1,044회), jIsc(897회), altmetrics(652회)순으로 높은 수치를 나타냈으며 다운로드의 경우에는 article processing charges(140회), altmetric(122회), open access(90회) 순으로 높은 수치를 나타냈다. 반면 library, research data, citation 등의 영역은 상대적으로 저조한 수치를 나타내, 오픈 엑세스 및 계량정보학 분야의 활용도가 다른 영역에 비해 높은 것으로 해석할 수 있겠다.

마지막으로 데이터 공개 수준과 활용 정도간의 상관성을 살펴본 결과를 제시해 보도록 한다. 〈표 9〉와 같이 분석 대상 데이터는 1단계에서 3단계까지의 공개 수준을 보이는데, 뷰 횟수는 단계가 상승할 때마다 약간씩 증가하는 경향을 보이지만, 다운로드 횟수는 그렇지 않은 것으로 나타났다. 스피어만 상관분석 결과

〈표 7〉 유형별 활용 평균 및 크루스칼 왈리스 검정 결과

type	뷰횟수 평균	다운로드 평균
code	227.8	12.9
dataset	1,092.6	148.1
figure	319.5	22.5
fileset	482.3	58.6
media	142.2	7.0
paper	804.2	160.3
poster	459.9	64.2
presentation	285.4	37.3
thesis	238.4	67.6
검정값	뷰	다운로드
카이제곱	182.157	282.140
자유도	8	8
근사 유의확률	.000	.000

〈표 8〉 주제분야별 활용 평균 및 크루스칼 왈리스 검정 결과

주제	뷰횟수 평균	다운로드 평균
open access	621.4	90.6
research data	327.0	46.7
altmetric	652.3	122.7
scholarly communication	484.8	73.7
article processing charges	1044.0	140.1
bibliometrics	337.5	79.3
libraries	228.3	51.8
jisc	897.4	90.1
open science	335.2	38.8
citation	302.4	57.3
검정값	뷰	다운로드
카이제곱	63.89747	50.61875
자유도	9	9
근사유의확률	.001	.000

〈표 9〉 데이터공개수준과 활용 정도 간의 스피어만 상관 분석 결과

공개수준		뷰	다운로드		뷰	다운로드
1단계	평균	502.9	78.8	Spearman의 rho	.160(**) 유의확률(양측).000	.010 유의확률(양측).793
2단계		644.4	72.4			
3단계		807.7	94.6			

에서도 데이터의 공개 수준이 발전될수록 뷰횟수(r=.160)는 약간 증가하는 추세를 보였으나 다운로드(r=.010)의 경우는 상관성이 전혀 없는 것으로 나타났다. 다시말해 아직까지 데이터가 편집과 재사용이 가능한 2-3단계 상태로 공개되어 있다고 할지라도 재사용도가 크게 높아지지 않는다는 것이다. 그 원인은 앞에서도 언급한 바와 같이 데이터가 가공이 가능한 수준으로 공개되어 있다고 할지라도 공개된 데이터의 변수와 조사방법론에 대한 충분한 설명이 부족해 해석과 재활용이 자유롭지 않은 경우가 존재할 뿐 아니라 원데이터를 재활용해 새로운 연구 성과를 창출할 수 있다는 인식도 저조하기 때문인 것으로 추정해 볼 수 있을 것이다.

5. 결론 및 제언

Figshare를 통해 공개되고 있는 문헌정보학 분야 연구데이터의 현황과 공개 수준 및 공유 정도를 파악한 결과를 종합하여 논의하면 다음과 같다.

첫 번째, 다양한 데이터 유형 중 dataset과 paper 형태의 연구데이터가 Figshare에 가장 많이 구축되어 있으며, 주제 영역에 있어서는 open access, RDM, altmetrics의 비중이 높은 것으로 나타났다. 연구데이터는 실질적인 재활용을 위해 다운로드된 경우보다 뷰된 경우가 7배 정도 많았으며 뷰, 다운로드, 트윗된 횟수 간에는 강한 상관성이 존재하였다. 그러나 다수 공유된 연구데이터는 소수에 이를 뿐이고 대부분의 데이터는 공유도가 낮아 긴꼬리 모양의 그래프를 나타내고 있었다.

두 번째, 연구데이터의 공개 수준을 살펴본 결과 70%에 가까운 대부분의 연구데이터가 pdf와 같이 편집과 가공이 원활하지 않은 형태로 공개되고 있었다. 응용소프트웨어를 통해 가공한 2단계 수준의 공개 데이터와 오픈 포맷을 채택한 3단계 공개 수준의 데이터는 소수인 것으로 나타났다.

세 번째, 교차 분석과 대응일치 분석을 통해 데이터의 유형과 파일 형식간의 관계 분석을 수행한 결과 media의 경우는 mp3-4, figure, poster의 경우는 png, jpeg, tiff와 같은 이미지 화일간의 연관성이 두드러지게 나타났다. 그러나 dataset, fileset, paper 등의 데이터는 pdf, csv, doc, zip 등 다양한 종류의 데이터 파일 형식과 관련을 맺고 있는 것으로 나타났다. 한편, 데이터의 유형과 주제 간의 관계 분석 결과, dataset 형태로 가장 많이 제공되고 있는 주제 분야는 article processing charge를 비롯한 오픈엑세스 영역으로 나타났으며, 재활용의 효용이 높을 것이라 여겨지는 데이터 사이언스와 계량정보 분야는 의외로 dataset보다는 fileset이나 paper 형태로 연구데이터가 개방되고 있는 것으로 나타났다.

네 번째, 연구데이터의 특성과 활용 정도 간의 관계 분석 결과, 주제에 있어서는 APC를 비롯한 open access 영역과 altmetrics와 관련된 연구데이터가 가장 많이 공유되고 있는 것으로 나타났다. 데이터 유형에 있어서는 dataset보다 paper가 더욱 다운로드 횟수가 높은 것으로 나타나, 원데이터보다 데이터 분석 결과가 해석된 paper, 그리고 상업적 유통 채널로 확보하기 어려운 다양한 종류의 회색문헌을 인용하거나

참고하는 경우가 더욱 많은 것으로 추정되었다. 또한 이러한 추정은 연구데이터의 공개 수준과 활용 정도간의 상관성 분석 결과로도 설명되었다. 데이터 공개 수준과 다운로드 횟수 간에 아무런 상관성이 존재하지 않는 것으로 나타나, 아직까지 기계적 가공이 가능한 형태로 데이터가 공개되어 있다고 할지라도 원데이터를 재활용해 새로운 연구 성과를 창출할 수 있다는 인식으로 활발하게 이어지고 있지 않다고 말할 수 있겠다.

문헌정보학 분야에서는 다양한 방법으로 데이터를 수집해 도서관 및 정보 관련 현상을 설명하는 연구가 이루어지고 있다. 그러나 아직까지는 연구 결과에 대한 참고를 목적으로 부산물들이 공유되고 있으며 원데이터 자체를 활용한 새로운 연구 성과 창출로는 활발하게 이어지지 않고 있는 것으로 보인다. 원데이터의 재활용을 높이기 위해서는 소프트웨어에 독립적인 형식을 갖춘 기계 가독형 데이터가 공개되어야 할 것이며, 더욱이 변수와 방법론에 대한 표준화된 설명을 통해 데이터를 재활용하고자 하는 연구자들을 지원할 수 있어야 할 것이다. 최근 전 세계적으로 공공데이터의 개방과 재활용을 도모하는 데이터 민주화가 추구되면서 데이터를 기반으로 한 유용한 인프라를 구축하기 위하여 다양한 노력이 이루어지고 있다. 우리나라에서도 공공데이터의 제공 및 이용 활성화에 관한 법률(법률 제14839호)을 시행함으로써 공공기관이 데이터를 개방할 뿐 아니라, 데이터가 재활용될 수 있는 품질 수준을 유지할 수 있도록 정책적으로 유도하고 있다. 이러한 시류에 따라 학술커뮤니티에서도 연구데이터의 개방과 공유를 위한 노력을 체계적으로 시작해야 할 것이다. 연구데이터 공개와 체계적인 관리를 위해서는 관련 정책의 개발, 데이터 레포지토리의 구축, 연구자들의 인식 개선 등 선제되어야 할 과제가 산적하지만, 앞에서 시사하고 있는 바와 같이 공개된 데이터의 재활용성을 극대화할 수 있는 공개 표준이나 개방 지침의 개발이 병행되어야 할 것이다. 이를 위해 먼저 각기 다른 연구 방법론을 가진 학문 분야별로 연구데이터의 특성을 분석하는 후속 연구가 이어지길 기대한다.

참 고 문 헌

강주연 (2017). 생명공학분야 연구데이터 관리 방안 연구. 석사학위논문, 전북대학교, 기록관리학과.

김상수 (2005). 이분형 자료를 이용한 수질평가. Journal of the Korean Data Analysis Society, 7(1), 151-158.

김운봉, 김용민, 양진옥 (2014, December 24). 유전체 빅데이터 연구 동향. Retrieved from http://m.bioin.or.kr/board.do?num=249060&bid=report&cmd=view

김은정, 남태우 (2012). 연구데이터 수집에 영향을 미치는 요인 분석. 정보관리학회지, 29(2), 27-44. https://doi.org/10.3743/kosim.2012.29.2.027

김지현 (2012). 대학 내 연구자들의 연구데이터 관리에 관한 연구. 한국도서관·정보학회지, 43(3), 433-455. https://doi.org/10.16981/kliss.43.3.201209.433

김지현 (2013). 국외 정부연구비지원기관의 연구데이터 관리정책 분석. 한국문헌정보학회지, 47(3), 251-274. https://doi.org/10.4275/kslis.2013.47.3.251

김지현, 정은경, 윤정원, 이재윤 (2017). 데이터 인용의 현황과 제언. 정보관리학회지, 34(1), 7-29. https://doi.org/10.3743/kosim.2017.34.1.007

류귀열 (2011). 초고속 무선 인터넷에서 폰형과 모뎀형 단말기의 이용장소에 따른 선호 컨텐츠에 관한 연구. 한국데이터정보과학회, 22(4), 701-716.

백인수 (2013). 오픈데이터 플랫폼과 국가데이터 전략방향. [대구]: 한국정보화진흥원.

신영란, 정연경 (2012). 국내 인문사회 연구데이터 아카이브의 개선방안에 관한 연구. 한국기록관리학회지, 12(3), 93-115.

심원식 (2016). 미국 대학도서관의 연구데이터 지원 서비스 사례 연구. 한국문헌정보학회지, 50(4), 311-332. https://doi.org/10.4275/kslis.2016.50.4.311

심원식, 안혜연, 변제연 (2015). 인문학 분야 연구데이터의 수집 및 활용성 증진을 위한 전략 연구. 한국문헌정보학회지, 49(3), 155-183. https://doi.org/10.4275/kslis.2015.49.3.155

조재인 (2016). Data Citation Index를 기반으로 한 연구데이터 인용에 관한 연구. 한국문헌정보학회지, 50(1), 189-207. https://doi.org/10.4275/kslis.2016.50.1.189

채희원 (2017). 대응일치분석을 이용한 4년제 대학졸업자들의 취업훈련기관별 노동시장 성과에 영향을 미치는 요인. 한국산학기술학회, 18(4), 235-241.

최형욱, 정은경 (2017). 사회학 분야의 연구데이터 특성과 지적구조 규명에 관한 연구. 정보관리학회지, 34(3), 109-124.

科學技術·學術審議會 學術分科會 學術情報委員會 (2015). 學術情報のオープン化の推進について. Retrieved from http://www.mext.go.jp/component/b_menu/shingi/toushin/_icsFiles/afieldfile/2015/10/06/1362565_1.pdf

篠田麻美 (2014). 研究データの共有と活用のために：研究データ同盟の分科會. カレントアウェアネス, 253. Retrieved from http://current.ndl.go.jp/e1531

Aleixandre-Benavent, R., Moreno-Solano, L. M., Sapena, A. F., & Sánchez Pérez, E. A. (2016). Correlation between impact factor and public availability of published research data in information science and library science journals. Scientometrics, 107(1), 1-13. https://doi.org/10.1007/s11192-016-1868-7

Borrego, Á., & Garcia, F. (2013). Provision of supplementary materials in library and information science scholarly journals. Aslib Proceedings, 65(5), 503-514.

https://doi.org/10.1108/ap-10-2012-0083

CENDI (2017). Implementation of public access programs in federal agencies. Retrieved from https://cendi.gov/projects/Public_Access_Plans_US_Fed_Agencies.html

Crawford, W. (2015). Open access journals 2014, DOAJ subset. figshare. Retrieved from https://doi.org/10.6084/m9.figshare.1299451.v4

DataCite (2012). Business models principles. Retrieved from https://www.datacite.org/documents/Business_Models_Principles_v1.0.pdf

Department for Business, Innovation & Skills Prime Minister's Office (2013, June 12). G8 Science ministers statement: London. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

Digital-science (2015). GRID release 2015, Figshare [Data file]. Retrieved from https://doi.org/10.6084/m9.Figshare.2010108.v6

Hahnel, M. (2012). Reproducibility of research - A new standard. Retrieved from https://Figshare.com/blog/Reproducibility_of_research_-_A_new_standard/45

Halfaker, A., & Taraborelli, D. (2015). Scholarly article citations in Wikipedia, Figshare. Retrieved from https://doi.org/10.6084/m9.Figshare.1299540.v8

Japan Science & Technology Agency (2017). Implementation guidelines: JST policy on open access to research publications and research data management. Retrieved from http://www.jst.go.jp/EN/about/openscience/guideline_openscience_en.pdf

Kiley, R. (2014). Wellcome Trust APC spend 2012-13: data file. figshare. Retrieved from https://doi.org/10.6084/m9.figshare.963054.v1

Kramer, B., & Bosman, J. (2015). 101 Innovations in scholarly communication. Poster presented at FORCE 2015. Retrieved from https://www.force11.org/meetings/force2015

Lawson, S., Meghreblian, B., & Brook, M. (2015). Journal subscription costs - FOIs to UK universities, Figshare. Retrieved from https://doi.org/10.6084/m9.Figshare.1186832.v23

Organizatioon for Economic Co-operation and Development (2007). OECD principles and guidelines for access to research data from public funding. Retrieved from http://www.oecd.org/sti/sci-tech/38500813.pdf

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. Scientometrics, 107(2), 723-744. https://doi.org/10.1007/s11192-016-1887-4

Research Councils UK (2015). RCUK common principles on data policy. Retrieved from

http://www.rcuk.ac.uk/documents/documents/rcukcommonprinciplesondatapolicy-pdf/

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Government Information Quarterly, 30, S19-S31. https://doi.org/10.1016/j.giq.2012.06.011

The Office of Science and Technology Policy (2013). Increasing access to the results of federally funded scientific research. Paper presented at the meeting of the Executive Office of the President, Washington, DC. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Thelwall, M., & Kousha, K. (2016). Figshare: A universal repository for academic resource sharing? Online Information Review, 40(3), 333-346. https://doi.org/10.1108/oir-06-2015-0190

Torres-Salinas, D., Martín-Martín, A., & 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), e036. https://doi.org/10.3989/redc.2014.1.1114

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Beak, In-soo (2013). Open data platform and national data strategy direction. [Daegu]: National Information Society Agency.

Chae, Hee Won (2017). The determinants of labor market outcomes in four-year graduates through principal component analysis and correspondence analysis by training institution. Journal of the Korea Academia-Industrial Cooperation Society, 18(4), 235-241.

Cho, Jane (2016). Study about research data citation based on DCI (Data Citation Index). Journal of the Korean Society for Library and Information Science, 50(1), 189-207. https://doi.org/10.4275/kslis.2016.50.1.189

Choi, Hyung Wook & Chung, EunKyung (2017). An investigation on characteristics and intellectual structure of sociology by analyzing cited data. Journal of the Korean Society for Information Management, 34(3), 109-124.

Kang, Joo-yeon (2017). A study on the methods for biotechnology research data management. Master's thesis, Graduate School of Chonbuk National University, Department of Records & Archives Management.

Kim, Eun-Jeong, & Nam, Tae-Woo (2012). Factor analysis of effects on research data collection. Journal of the Korean Society for Information Management, 29(2), 27-44. https://doi.org/10.3743/kosim.2012.29.2.027

Kim, Jihyun (2012). A study on university researchers' data management practices. Journal of the Korean Library and Information Science Society, 43(3), 433-455. https://doi.org/10.16981/kliss.43.3.201209.433

Kim, Jihyun (2013). An analysis of data management policies of governmental funding agencies in the U.S., the U.K., Canada and Australia. Journal of the Korean Society for Library and Information Science, 47(3), 251-274. https://doi.org/10.4275/kslis.2013.47.3.251

Kim, Jihyun, Chung, EunKyung, Yoon, JungWon & Lee, Jae Yun (2017). The current state and recommendations for data citation. Journal of the Korean Society for Information Management, 34(1), 7-29. https://doi.org/10.3743/kosim.2017.34.1.007

Kim, Sang Soo (2005). Possibility of evaluation water quality using binary data. Journal of the Korean Data Analysis Society, 7(1), 151-158.

Kim, Unbong, Kim, Yongmin, & Yang, Jinok (2014, December 24). Trends in genetic data research. Retrieved from http://m.bioin.or.kr/board.do?num=249060&bid=report&cmd=view

Ryu, Kuiyeol (2011). A study on preferable contents depending on regions and terminal types for high speed mobile internet. Journal of the Korean Data & Information Science Society, 22(4), 701-716.

Shim, Wonsik (2016). A case study of U.S. academic libraries' research data support services. Journal of the Korean Society for Library and Information Science, 50(4), 311-332. https://doi.org/10.4275/kslis.2016.50.4.311

Shim, Wonsik, Ahn, Hye-yeon, & Byun, Jeayeon (2015). Strategies for improving the collection and use of research data in the humanities. Journal of the Korean Society for Library and Information Science, 49(3), 155-183. https://doi.org/10.4275/kslis.2015.49.3.155

Shin, Young-Ran, & Chung, Yeon-Kyoung (2012). A study on the improvement plans of the humanities and social sciences research data archives in Korea. Journal of Records Management & Archives Society of Korea, 12(3), 93-115.