사회학 분야의 연구데이터 특성과 지적구조 규명에 관한 연구

An Investigation on Characteristics and Intellectual Structure of Sociology by Analyzing Cited Data

최형욱 (Hyung Wook Choi)*
정은경 (EunKyung Chung)**

초 록

여러 학문 분야에서 데이터의 공유와 재이용에 관한 관심이 증가하고 있다. 실제로 다른 연구자의 데이터를 다시 연구에 사용하고 인용을 부여하는 관행이 서서히 자리를 잡아가고 있다. 이러한 변화를 반영하여 톰슨로이터는 Data Citation Index(DCI)라는 데이터인용 색인 데이터베이스 서비스를 2012년부터 제공하기 시작하였다. DCI는 모든 학문의 전 영역에서 데이터의 인용 현황을 저널의 논문과 유사하게 집계한다. 본 연구에서는 데이터인용이 활발한 사회학 분야의 인용된 연구데이터를 분석하여 해당 분야의 특성과 지적구조를 규명하고자 하였다. 이를 위해 논문 인용을 기반으로 한 사회학 분야의 지적구조와 비교하였으며, 사회학 분야의 연구데이터의 특성과 고유한 지적구조를 살펴보고자 하였다. 분석을 위한 데이터는 두 종류로 수집하였다. 첫째는 DCI에서 'Sociology'로 주제 검색을 수행하여 총 8,365건의 인용된 데이터를 수집하였다. 둘째로, 논문 인용 분석과의 비교를 위해서 Web of Science에서 'Sociology'로 주제 검색을 수행하여 총 12,132건의 데이터를 수집하였다. 이 두 데이터를 활용하여 저자키워드 동시출현단어 분석을 수행한 결과, 데이터를 기반으로 한 사회학 분야는 2영역 15군집으로 구성된 반면, 논문을 기반으로 한 사회학 분야는 3영역 17군집으로 나타났다. 내용적인 특성을 살펴보면, 전통적으로 사회학의 지적구조를 나타낸다고 볼 수 있는 논문 기반 사회학과 달리 사회학 분야의 연구데이터는 의학 분야와의 활발한 접목을 찾아볼 수 있으며, 그 중에서도 공중보건과 심리학이 중심 영역인 것으로 나타났다.

ABSTRACT

Through a wide variety of disciplines, practices on data access and re-use have been increased recently. In fact, there has been an emerging phenomenon that researchers tend to use the data sets produced by other researchers and give scholarly credit as citation. With respect to this practice, in 2012, Thomson Reuters launched Data Citation Index (DCI). With the DCI, citation to research data published by researchers are collected and analyzed in a similar way for citation to journal articles. The purpose of this study is to identify the characteristics and intellectual structure of sociology field based on research data, which is one of actively data-citing fields. To accomplish this purpose, two data sets were collected and analyzed. First, from DCI, a total of 8,365 data were collected in the field of sociology. Second, a total of 12,132 data were collected from Web of Science with a topic search with 'Sociology'. As a result of the co-word analysis of author provided-keywords for both data sets, the intellectual structure of research data-based sociology was composed of two areas and 15 clusters and that of article-based sociology was composed with three areas and 17 clusters. More importantly, medical science area was found to be actively studied in research data-based sociology and public health and psychology are identified to be central areas from data citation.

키워드: 사회학, 동시출현단어분석, DCI, 지적구조, 네트워크 분석
sociology, co-word analysis, DCI, intellectual structure, network analysis

* 이화여자대학교 일반대학원 문헌정보학과 석사(shinywook92@gmail.com) (제1저자)
** 이화여자대학교 사회과학대학 문헌정보학과 부교수(echung@ewha.ac.kr) (교신저자)

■ 논문접수일자 : 2017년 8월 20일 ■ 최초심사일자 : 2017년 9월 2일 ■ 게재확정일자 : 2017년 9월 9일
■ 정보관리학회지, 34(3), 109-124, 2017. [http://dx.doi.org/10.3743/KOSIM.2017.34.3.109]

1. 서 론

최근 정보기술의 발전과 함께 오픈사이언스에 대한 관심이 증가함에 따라 연구 데이터에 대한 공유와 재이용이 여러 연구 분야에서 주목을 받고 있다. 데이터 공유는 연구의 효율성을 높이고 연구 과정의 투명성을 제공할 뿐만 아니라 다양한 학문 분야의 연구데이터를 통합하고 재해석하여 새로운 가치를 창출할 수 있도록 한다는 장점을 지니고 있다(Dalerba, Sahoo, Paik, Guo, Yothers, Song, … Hisamori, 2016; Parr & Cummings, 2005; Peng, 2011; Stodden, Leisch, & Peng, 2014). 특히 학문의 연구 과정에서 모든 유형의 지식을 공개하고 공유해야 한다(European Commission, 2015)는 개념의 오픈사이언스 시대에서 연구데이터의 공유는 새로운 지식 패러다임을 창출하기 위한 기반으로 여겨지고 있다.

데이터의 공유와 재이용을 촉진하는 여러 환경 중에서 연구기금기관의 역할은 특히 중요하다. 현재 미국의 주요 기금 지원 기관인 국립암센터나 미국국립보건원, 미국국립과학재단의 경우, 연구자들에게 데이터세트와 컴퓨터 코드와 같은 데이터 공유를 정책적으로 장려하고 있으며, 연구 계획서와 함께 데이터 공유를 위한 데이터 관리 계획을 반드시 함께 제출할 것을 요구하고 있다(Diekema, Wesolek, & Walters, 2014; National Cancer Institute, 2006; National Institutes of Health, 2003). Ross, Tse, Zarin, Xu, Zhou, Krumholz(2012)는 미국의학연구소의 기금을 받아 수행되는 의학 임상실험 중 3분의 1 규모의 데이터가 실험이 끝난 후 4년 동안 공개되지 않으며, 대부분의 데이터가 분석되어 지지 않는다는 것을 지적하였다. 또한 Lo(2015)는 임상실험 데이터를 공개하도록 하는 정책을 마련하여 연구의 결과를 출판하고 학술적 표준을 적용하여 연구 업적을 평가하는 의학 저널들을 대상으로 시행하여야 한다고 주장하였다. 미국과 마찬가지로 유럽 국가 역시 국가 연구의 경우, 데이터 공유를 하지 않는 경우 기금을 제한하는 등 데이터 개방을 정책적으로 시행하는 추세이다(Spires-Jones, Poirazi, & Grubb, 2016).

연구자의 관점에서 데이터 공유와 재이용이 장려되는 상황은 중요한 논의점 중 하나이며, 데이터 출판과 이에 상응하는 적절한 인정과 저작자 표시는 연구자에게 상당한 동기유발로 작용할 수 있다. 때문에 최근 연구과정에서 생산되는 데이터를 연구자의 학술적 성과물로 인식하고, 더 나아가 데이터인용이라는 방식을 통해 적절한 학술적 인정을 부여해야 한다는 인식이 확산되고 있으며, 연구데이터의 체계적인 관리와 데이터인용을 위한 표준에 대한 필요성 역시 대두되고 있다. 이러한 상황에서 톰슨로이터(Thomson Reuters)는 2012년 10월 Data Citation Index(이하 DCI) 서비스를 제공하고 시작하였다. DCI는 연구데이터를 수집하여 개방 및 공유하는 데이터 레포지토리를 선정하고 색인하여 각 기관의 피인용도에 따른 영향력 측정이 가능한 데이터를 함께 제공하고 있다.

본 연구에서는 데이터인용 색인 데이터베이스인 DCI 데이터를 기반으로 학문 분야의 특성과 지적구조를 규명하고, 이를 논문 기반 지적구조 분석결과와 비교하였다. 이를 통하여 데이터 공유와 재이용을 기반으로 새롭게 나타나는 학문의 변화를 조명하고자 한다. 분석 대상은 활발한 데이터 인용이 이루어지고 있는 사

회학(Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015) 분야로 선정하였다. 이러한 연구목적을 성취하기 위해서 구체적으로 세 가지 연구목표를 설정하였다. 첫째, 데이터 공유와 재이용이 이루이지는 사회학 분야의 특성을 규명하고자 사회학 분야에서 인용된 연구데이터의 특성 분석을 수행하였다. 둘째, 연구데이터의 저자키워드 동시출현단어 분석을 활용하여 지적구조를 제시하고자 하였다. 셋째, 논문 기반 사회학 지적구조와 데이터 기반 사회학 지적구조를 비교하여 데이터 기반 사회학에서 찾아볼 수 있는 특징을 밝히고자 하였다.

2. 관련 연구

데이터인용에 대한 연구는 데이터 공유와 개방의 필요성이 확산되면서 지속적으로 이루어졌다. 특히 미국국립보건원과 미국국립과학재단이 국가의 지원을 받은 연구는 데이터를 반드시 개방하여 다른 연구자와 공유해야 한다는 공식 성명서를 발표하면서(Simberloff, Barish, Droegemeier, Etter, Fedoroff, Ford, … White Jr., 2005), 미국 내의 데이터 공유와 개방은 더욱 활발하게 이루어지고 있다.

우선 Mooney와 Newton(2012)은 Data Citation Adequacy Index(이하 DCAI)을 제시하였다. 세 학문 범주에서 데이터인용에 나타난 연구자 행위, 학문분야별 인용을 위한 정보원과 스타일 양식, 데이터 발행처, 학술 논문 등의 내용분석을 통하여 본문과 인용에 나타난 특성을 분석하였다. 인문학, 사회과학, 과학 분야는 절반 이상의 학술지에서 데이터인용을 위한 양식과 표준을 제시하고 있지만, 대부분의 학술 논문에서 2차 분석 연구에 쓰인 데이터에 대한 적절한 인용을 포함하고 있지 않음을 규명하였다. 다양한 데이터 유형과 존재하는 기준들에 대한 부족한 인식을 데이터인용이 활발하지 못한 원인으로 지적하며, 도서관 사서들과 정보 전문가들이 다양한 분야에 걸친 데이터인용을 촉진시키는 데 고무적인 역할을 담당해야 한다고 제시하였다. 또한 Belter(2014)는 국립 해양학 데이터 센터(National Oceanographic Data Center)에 소장된 데이터 인용 빈도수를 활용하여 서지계량적 영향을 제시하고자 하였다. 연구 결과에 따르면 데이터는 상당한 인용 비율을 가지고 있었다. 특히, 특정 데이터는 저널에서 발행된 학술 논문의 인용률 보다 높다는 것을 밝혔다. 또한 각 데이터의 공식적 인용 형식이 제안되어 있음에도 불구하고, 과학 출판물에서 이러한 데이터세트를 인용하는 방법에 일관성이 없는 것으로 나타났다. 이러한 연구 결과는 연구자가 연구데이터 생산하여 적절하게 출판하는 것이 필요하며, 데이터인용을 통한 개인과 기관의 영향력 평가에 대한 중요성을 제시하였다.

한편 최근의 연구 중에서는 DCI를 분석하여 연구결과를 제시하는 경향을 찾아볼 수 있다. Robinson-García, Jiménez-Contreras, Torres-Salinas(2015)는 2013년 5월과 6월에 걸쳐 DCI에 색인되어 있는 데이터를 추출하여 연도별, 주제별, 데이터 유형별, 리포지토리 분석을 수행하였다. 그 결과, 전체 데이터 중 데이터세트 유형의 비중이 컸으며, 전체 데이터 중 대부분을 차지하는 88.1%의 데이터가 실제로 인용되지 않은 것으로 나타났다. 데이터세트 인용은 대부분 과학과 공학/기술 분야로 나타났으며,

데이터스터디의 경우 사회과학과 인문예술 분야에서 인용이 활발하게 이루어졌다고 밝혔다. 전체적인 데이터의 규모는 과학 분야에 치중되어 있었으나, 데이터스터디 유형의 경우 전체 인용의 30.8%가 사회학 분야에서 이루어진 것으로 나타났다. 또한, 데이터인용이라는 개념이 점차 다양한 학문 분야로 확대됨에 따라 다양한 데이터 유형에 대한 연구에 대한 필요성을 제시하였다. 또한 조재인(2016)은 DCI로부터 데이터를 추출하여 전체적인 데이터의 규모와 유형, 국가별 분포, 주제별 분포, 주요 데이터 리포지토리 현황을 살펴보았다. 또한, 피인용수가 높은 데이터 500건에 대해 데이터의 유형과 주제가 인용도에 미치는 영향을 살펴보고, 사회적 영향력을 분석하였다. 분석을 통해 데이터의 규모는 유전학과 생명공학 분야가 가장 큰 것으로 나타났으나, 인용이 많이 되는 분야는 사회과학분야로 확인되었다. 또한 데이터의 유형 역시 그 규모는 데이터세트가 더 크지만, 인용이 많이 되는 유형은 데이터스터디라고 밝혔다. 높은 인용도에 영향을 주는 요인으로는 데이터스터디라는 데이터 유형과 사회과학분야라는 주제를 꼽았다. 사회적 영향력 분석에서는 대부분의 데이터가 그 영향력이 미비한 것으로 나타났다. 전 세계적으로 데이터 공유와 인용에 대한 인식이 확대됨에 따라 우리나라 역시 그에 발맞춘 연구데이터 관리에 대한 교육과 각종 기반을 마련해야 한다고 지적하였다.

특정 지역의 데이터 인용 현황을 분석한 연구는 Onyancha(2016)에 의해서 수행되었다. 연구기간은 2009년부터 2014년으로 한정하여 DCI의 데이터를 추출하여 사하라 사막 이남의 아프리카(sub-Saharan Africa, 이하 SSA) 지역의 연구데이터 공유 현황에 대해 살펴보았다. 이를 위해 국가별, 기관별, 주제 분야별, 출판연도별, 문서 유형별 그리고 인용 횟수에 대한 분석을 수행하였다. SSA 지역의 경우, 국민의 다수가 농업에 종사한다는 특징으로 인해 데이터의 주제 분야 분석에서 농업이 나타나는 특이점을 보였다. 대부분의 데이터베이스가 북미와 유럽 국가들을 중심으로 한 선진국에서 개발되었음을 지적하면서 아프리카 지역 자체적인 데이터베이스 설립의 필요성에 대해 지적하였다.

이러한 관련 연구는 대체적으로 데이터 인용의 전반적인 현황을 제시하였다. 이에 따라 본 연구는 기존의 인용된 논문의 분석을 기반으로 한 지적구조와 인용된 데이터를 기반으로 한 지적구조를 비교하여 데이터 인용을 기반으로 한 지적구조가 가지는 고유한 특성을 규명하고자 하였다.

3. 연구 방법

3.1 데이터 수집과 키워드 선정

연구를 위해 데이터 인용이 활발하게 이루어지고 있는 사회과학 분야(조재인, 2016; Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015) 중에서 인용 비율이 높게 나타난 사회학 분야를 분석 분야로 선정하여 두 가지 데이터를 수집하였다. 첫 번째 데이터 수집을 위해 DCI에서 'Sociology'로 주제 검색하였으며, 2017년 4월 19일 기준 총 8,365건을 수집하였다. 두 번째 데이터를 수집하기 위해 논문 인용색인 데이터베이스인 WoS에서 문서 유형을 'Article'로 제한하고 'Sociology'로 주제 검색을 수행하였다. 그 결과, 2017년 4월 10일 기준으로 총 12,132건의

데이터를 수집하였다.

Bibexcel[1] 프로그램을 사용하여 저자 키워드(DE)를 추출하였다. 데이터를 기반으로 한 사회학 분야의 전체 8,365건의 데이터에서 총 151,629개의 저자 키워드를 얻을 수 있었다. 또한 논문을 기반으로 한 사회학 분야의 전체 12,132건의 데이터에서 총 49,805개의 저자 키워드를 추출하였다. 이후 동시출현단어 네트워크 분석을 통한 지적구조를 규명하기 위해서 저자 키워드의 빈도수를 각각 300회와 60회로 제한하여 다음 〈표 1〉과 〈표 2〉와 같이 각각 78개와 61개의 최종 키워드를 선정하였다.

〈표 1〉 사회학 분야의 연구데이터 저자 키워드 상위 78개

번호	키워드	출현횟수	번호	키워드	출현횟수
1	Sociology	8,036	40	Health Economics	520
2	Medicine	4,789	41	Public And Occupational Health	519
3	Biological Sciences	4,256	42	Social Networks	517
4	Psychology	3,704	43	Evolutionary Biology	502
5	Mental Health	2,266	44	Information And Computing Sciences	499
6	Behavior	1,946	45	Psychological Stress	493
7	Public Health	1,878	46	Health Care	460
8	Epidemiology	1,664	47	Human Performance	456
9	Non-Clinical Medicine	1,589	48	Computational Biology	450
10	Neuroscience	1,585	49	Linguistics	450
11	Science Policy	1,577	50	Neuroimaging	434
12	Clinical Research Design	1,563	51	Earth And Environmental Sciences	433
13	Cognitive Psychology	1,322	52	Viral Diseases	416
14	Health Care Policy	1,280	53	Engineering	408
15	Economics	1,150	54	Statistics	408
16	Behavioral And Social Aspects Of Health	1,118	55	Clinical Psychology	400
17	Mathematics	918	56	Survey Methods	400
18	Experimental Psychology	854	57	Preventive Medicine	391
19	Anthropology	848	58	Applied Mathematics	389
20	Psychiatry	804	59	Nutrition	388
21	Socioeconomic Aspects Of Health	788	60	Research Design	388
22	Infectious Diseases	785	61	Sexual And Gender Issues	384
23	Cognitive Neuroscience	780	62	Child Health	377
24	Social Psychology	747	63	Diagnostic Medicine	367
25	Survey Research	702	64	Hiv	363
26	Population Biology	687	65	Psychometrics	343
27	Global Health	666	66	Human Relations	341
28	Neurology	639	67	Cognition	334
29	Social Research	614	68	Developmental Psychology	328
30	Anatomy And Physiology	607	69	Health Services Research	327
31	Demography	604	70	Personality	327
32	Ecology	599	71	Information Technology	323
33	Geography	592	72	Learning	323
34	Neuropsychology	567	73	Regression	321
35	Emotions	559	74	Political Science	320
36	Social Epidemiology	553	75	Human Geography	318
37	Communications	544	76	Statistical Methods	315
38	Pediatrics	533	77	Science Education	310
39	Sensory Perception	531	78	Mood Disorders	302

1) http://homepage.univie.ac.at/juan.gorraiz/bibexcel/

〈표 2〉 사회학 분야의 논문 저자 키워드 상위 61개

번호	키워드	출현횟수	번호	키워드	출현횟수
1	Sociology	1,216	32	Epistemology	86
2	Gender	254	33	Habitus	83
3	Culture	231	34	Knowledge	82
4	Bourdieu	191	35	Qualitative Research	79
5	Identity	179	36	Cultural Sociology	78
6	Economic Sociology	162	37	Higher Education	78
7	Globalization	145	38	Environmental Sociology	76
8	Sociology Of Knowledge	141	39	Embodiment	75
9	Ethnography	136	40	Neoliberalism	75
10	Religion	126	41	Violence	74
11	Methodology	116	42	History	71
12	Political Sociology	115	43	Actor-Network Theory	69
13	Education	113	44	Consumption	68
14	Power	113	45	Discourse	68
15	Theory	112	46	Innovation	67
16	Social Networks	109	47	Political Economy	65
17	Social Theory	109	48	Emotions	64
18	Ethics	105	49	Children	63
19	Social Capital	104	50	Markets	63
20	Inequality	102	51	Media	63
21	Class	101	52	Ethnicity	62
22	Race	101	53	Modernity	62
23	Sociology Of Science	101	54	Social Class	62
24	Politics	100	55	Sustainability	62
25	Risk	98	56	Trust	62
26	Historical Sociology	97	57	Family	61
27	Social Movements	92	58	Technology	61
28	Reflexivity	90	59	Climate Change	60
29	Health	89	60	Community	60
30	Public Sociology	89	61	Medical Sociology	60
31	Agency	87			

3.2 분석 과정

분석 대상 키워드를 선정한 후, 본 연구의 구체적인 목적을 위해 크게 두 단계의 분석을 수행하였다. 첫 번째 단계는 기술 분석(descriptive analysis)으로, 데이터 기반 사회학 분야 데이터에 대한 기초적인 분석을 수행하였다. 두 번째는 지적구조를 분석하여 비교하기 위한 단계로, 동시출현단어 분석을 수행하였으며 분석에는 이재윤이 개발한 COOC ver. 0.4를 활용하였다. 이를 통해 빈도수를 그대로 적용한 동시출현행렬, 코사인 유사도 계수로 정규화한 행렬, 피어슨 상관계수에 의한 2차 연관성 행렬을 구하였다. 네트워크와 세부 영역의 클러스터링을 구현하기 위해 이재윤의 WNET ver 0.4.1 프로그램을 사용하였으며 작성된 행렬을 네트워크로 시각화 하기 위해 NodeXL[2] 프로그램을 사용하였다.

2) https://nodexl.codeplex.com/

4. 분석 결과

4.1 사회학 분야 연구데이터의 특성

사회학 분야의 연구데이터에 대한 특성을 확인하기 위해 데이터 유형, 발행연도, 피인용 빈도를 〈표 3〉과 같이 살펴보았다. 총 8,365건의 데이터를 분석한 결과, 데이터 유형은 데이터스터디, 데이터세트, 소프트웨어, 레포지터리 순으로 나타났다. 이는 사회과학 분야에서 데이터스터디 비율이 전체 약 86%로 그 비율이 높다는 선행연구와도 일치하는 결과이다(조재인, 2016; Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015). 소프트웨어 유형인 데이터의 경우, 총 36건 모두가 코드이다. 이 중에서 34건은 Figshare에 소장되어 있으며, 2건은 Comprehensive R Archive Network에 소장되었다. 데이터의 발간연도는 1961년부터 2017년까지 분포되어 있었으며, 가장 많은 데이터가 발행된 연도는 2014년이었다. 〈표 3〉은 10년 단위로 데이터의 발행 분포를 제시한다. 피인용도 빈도의 경우, 전체의 약 86%는 적어도 1회 이상 인용된 것으로 확인되었다. 가장 많은 인용된 데이터는 "General Social Survey 1974"로 총 169번 인용되었으나, 약 14%의 데이터는 한 번도 인용되지 않은 것으로 나타났다. 이러한 결과는 DCI에 색인된 88.1%가 한 번도 인용되지 않았다는 선행연구(조재인, 2016; Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015)와 비교하였을 때 사회학 분야의 데이터가 인용되는 비율이 월등히 높다는 것을 알 수 있다. 또한 총 8,365건의 데이터를 레포지터리로 분석한 결과, 20개의 레포지터리에 분포되어 소장된 것으로 나타났다. 이 중에서 영국의 Figshare가 전체의 약 96%인 8,045건의 데이터를 소장하고 있었으며, 영국의 UK Data Archive가 104건, 네덜란드의 Data Archiving and Networked Services(DANS-KNAW)가 86건, 미국의 The Association of Religion Data Archives가 33개 순으로 그 뒤를 이었다.

〈표 3〉 사회학 분야의 연구데이터 특성

분석 항목	세부 내용	규모 (단위: 건)
데이터 유형	데이터스터디	7,215
	데이터세트	1,113
	소프트웨어	36
	레포지터리	1
발행연도	1960년대	6
	1970년대	21
	1980년대	19
	1990년대	27
	2000년대	168
	2010년대	8,146
피인용빈도	0회	1,157
	1회	7,151
	2회 이상 10회 미만	20
	10회 이상 100회 미만	32
	100회 이상	5
리포지토리	Figshare	8,045
	UK Data Archive	104
	Data Archiving and Networked Services	86
	The Association of Religion Data Archives	33
	ICPSR	20
	Australian Data Archive	15
	Odum Institute Data Archive	11
	Finnish Social Science Data Archive	10
	Zenodo	10

4.2 사회학 분야 연구데이터의 주제적 특성

사회학 분야 연구데이터의 주제적 특성을 파악하기 위하여 데이터의 주제분야, WoS 주제범주를 분석하였고, 그 내용은 〈표 4〉와 같다. 주제분야와 WoS 주제범주의 경우, 상위 10개의 분석 결과만을 표로 제시하였다. 주제분야는 연구자가 데이터가 속하는 학문 분야를 정하는 필드이다. 분석 결과, 총 32개의 주제분야에 분산되어 있었으며, 사회과학으로 분류되어 있는 사회학의 데이터임에도 불구하고 Science Technology Other Topics에 전체의 약 96%를 차지하는 8,029건이 분류되어 있었으며, 그 뒤를 이어 Sociology가 110건으로 나타났다. WoS 주제범주란 데이터가 색인될 때 톰슨로이터가 자체적으로 데이터를 분류하는 것으로 WoS의 저널 주제범주와 같은 학문 분류 체계를 적용하고 있다. 사회학 분야의 연구데이터를 분석한 결과, 총 38개의 범주가 확인되었다. 그 중 Multidisciplinary Sciences가 8,029건으로 가장 많은 부분을 차지하고 있었으며 Sociology가 110건, Humanities Multidisciplinary가 86건, Social Sciences Interdisciplinary가 55건으로 나타났다. 이러한 결과는 연구분야 분석의 결과와 비교하였을 때 그 주제적 분류 범주와 데이터의 건수가 대체적으로 일치함을 확인할 수 있었다.

〈표 4〉 사회학 분야의 연구데이터 주제적 특성

분석 항목	세부 내용	규모 (단위: 건)
주제분야	Science Technology Other Topics	8,029
	Sociology	110
	Arts Humanities Other Topics	86
	Social Science Other Topics	55
	Government Law	40
	Computer Science	36
	Religion	35
	Demography	26
	Education Educational Research	20
	Business Economics	17
WoS 주제범주	Multidisciplinary Science	8,029
	Sociology	110
	Humanities Multidisciplinary	86
	Social Science Interdisciplinary	55
	Political Science	40
	Computer Science Interdisciplinary Applications	36
	Religion	35
	Demography	26
	Education Educational Research	20
	Urban Studies	17

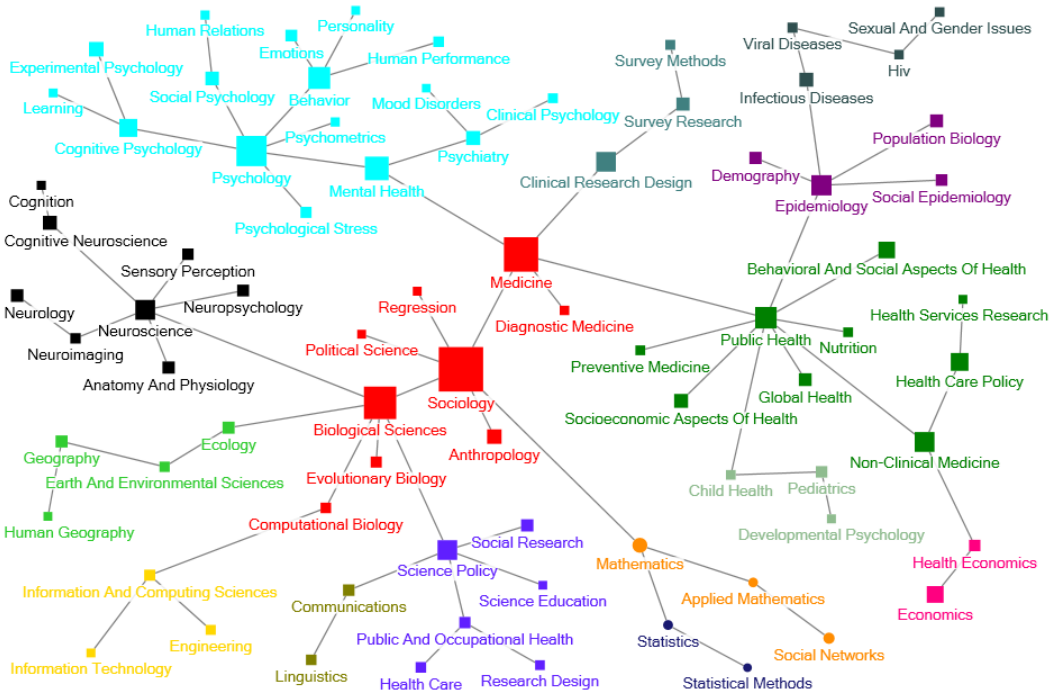
4.3 사회학 분야의 연구데이터 기반 지적 구조

사회학 분야의 연구데이터를 기반으로 한 지적구조를 파악하기 위해 저자키워드의 동시출현 단어를 대상으로 하였다. 산출된 코사인 유사도 계수로 정규화한 행렬에 $r=\infty$, $q=n-1$ 조건의 패스파인더 네트워크(PFNet) 알고리즘을 적용하였다. 또한, 병렬 최근접 이웃 클러스터링 알고리즘(PNNC)을 적용하여 패스파인더 네트워크상의 군집을 여러 개의 하위 네트워크로 분할하여 세부 주제를 구체적으로 파악하였다. 사회학 분야의 연구데이터 전체에서 추출한 저자 키워드 중 상위 78개 키워드를 대상으로 동시출현

단어의 네트워크 분석을 수행한 결과, 2개와 15개의 PNNC 최적 군집이 생성되었다. 세부 주제 분야를 파악하기 위해 이를 2개의 영역과 15개의 세부 군집으로 구분하였다. 〈그림 1〉은 2개의 영역을 노드의 모양으로, 15개의 세부 군집은 노드의 색깔로, 키워드 간의 빈도에 의한 연관도 가중치는 링크의 굵기로, 키워드별 출현횟수는 노드의 크기로 지정하여 표현하였다. 각 세부 군집의 대표 키워드는 군집별 키워드 중 빈도수가 가장 높게 나타난 키워드를 주제명으로 부여하였다. 〈표 5〉로 〈그림 1〉을 정리하였고, 영역별 세부 군집의 수와 세부 군집별 키워드의 개수를 명시하였다. 세부 군집별 대표 키워드로 나타난 군집명은 진하게 표시하였다.

〈표 5〉 연구데이터 기반 사회학 분야의 PNNC 세부 군집별 키워드

영역	세부 군집	세부 키워드	영역	세부 군집	세부 키워드
제1영역 (13 군집)	제1군집 (9개)	**Sociology**	제1영역 (13 군집)	제5군집 (8개)	**Neuroscience**
		Medicine			Cognitive Neuroscience
		Biological Sciences			Neurology
		Anthropology			Anatomy And Physiology
		Evolutionary Biology			Neuropsychology
		Computational Biology			Sensory Perception
		Diagnostic Medicine			Neuroimaging
		Regression			Cognition
		Political Science		제6군집 (6개)	**Science Policy**
	제2군집 (16개)	**Psychology**			Social Research
		Mental Health			Public And Occupational Health
		Behavior			Health Care
		Cognitive Psychology			Research Design
		Experimental Psychology			Science Education
		Psychiatry		제7군집 (3개)	**Clinical Research Design**
		Social Psychology			Survey Research
		Emotions			Survey Methods
		Psychological Stress		제8군집 (2개)	**Economics**
		Human Performance			Health Economics
		Clinical Psychology		제9군집 (4개)	**Infectious Diseases**
		Psychometrics			Viral Diseases
		Human Relations			Sexual And Gender Issues
		Personality			Hiv
		Learning		제10군집 (4개)	**Ecology**
		Mood Disorders			Geography
	제3군집 (9개)	**Public Health**			Earth And Environmental Sciences
		Non-Clinical Medicine			Human Geography
		Health Care Policy		제11군집 (2개)	**Communications**
		Behavioral And Social Aspects Of Health			Linguistics
		Socioeconomic Aspects Of Health		제12군집 (3개)	**Pediatrics**
		Global Health			Child Health
		Preventive Medicine			Developmental Psychology
		Nutrition		제13군집 (3개)	**Information And Computing Sciences**
		Health Services Research			Engineering
	제4군집 (4개)	**Epidemiology**			Information Technology
		Population Biology	제2영역 (2 군집)	제14군집 (3개)	**Mathematics**
		Demography			Social Networks
		Social Epidemiology			Applied Mathematics
				제15군집 (2개)	**Statistics**
					Statistical Methods



〈그림 1〉 연구데이터 기반 사회학 분야의 저자키워드 네트워크 분석(PFNet와 PNNC 세부 군집)

연구데이터를 기반으로 한 지적구조는 크게 두 영역으로 나뉘며, 첫 번째 영역은 총 13개의 군집을 포함하고 있다. 규모가 큰 순서로 보면, Psychology, Sociology, Public Health, Neuroscience, Science Policy, Epidemiology, Infectious Diseases, Ecology, Clinical Research Design, Pediatrics, Information and Computing Sciences, Economics, Communications이다. 두 번째 영역은 2개 군집을 포함하고 있으며 Mathematics와 Statistics이다. 그러나 〈그림 1〉에서 살펴볼 수 있는 바와 같이 Sociology 분야가 하위 키워드 수 측면에서는 Psychology 분야에 비해 적으나, 네트워크의 중심부에 위치하고 있어서 다른 분야와의 매개적 역할이라는 측면에서 그 중요도가 높다고 볼 수 있다.

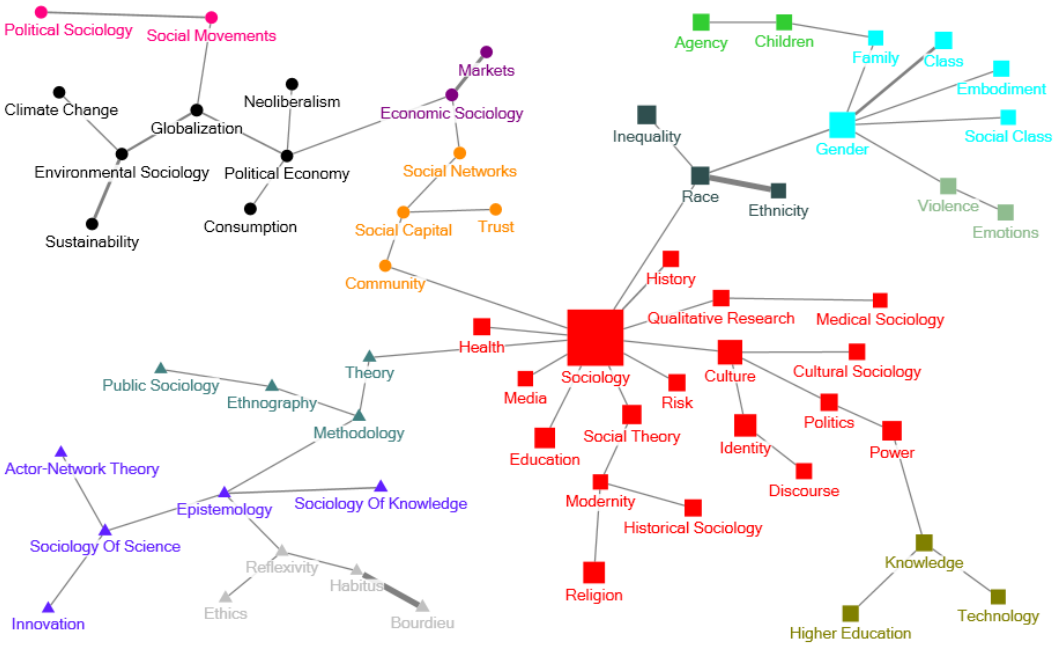
4.4 사회학 분야의 논문 기반 지적구조

사회학 분야 연구데이터의 지적구조와의 비교를 위해 사회학 분야 논문의 저자키워드 동시출현분석을 통해 지적구조를 규명하고자 하였다. 이를 위해 학문 분야의 세부 주제 영역의 개념들을 하위 군집으로 분할하여 네트워크로 시각화하여 파악할 수 있는 네트워크 분석 기법을 사용하였다. 연구데이터의 저자키워드 동시출현단어 분석과 마찬가지로 코사인 유사도 계수로 정규화한 행렬에 $r=\infty$, $q=n-1$ 조건의 패스파인더 네트워크(PFNet) 알고리즘을 적용하고, 다시 하위 군집으로 분할하기 위해 병렬 최근접 이웃 클러스터링 알고리즘(PNNC)을 적용하였다.

61개의 최종 키워드로 동시출현단어의 네트워크 분석을 수행한 결과, 3개 영역과 13개의 PNNC 최적 군집이 생성되었다. 〈그림 2〉는 3개의 영역을 노드의 모양으로, 13개의 세부 군집은 노드의 색깔로, 키워드의 출현횟수를 노드의 크기로, 키워드 간의 빈도에 의한 연관도 가중치는 링크의 굵기로 지정하여 표현하였다. 각 세부 군집의 대표 키워드는 군집별 키워드 중 빈도수가 가장 높게 나타난 키워드를 주제명으로 부여하였다. 〈표 6〉으로 〈그림 2〉를 정리하였고, 영역별 세부 군집의 수와 세부 군집별 키워드의 개수를 명시하였다. 세부 군집별 대표 키워드로 나타난 군집명은 진하게 표시하였다.

〈표 6〉 논문 기반 사회학 분야의 PNNC 세부 군집별 키워드

영역	군집	키워드	영역	군집	키워드
제1영역 (6 군집)	제1군집 (18개)	**Sociology**	제1영역 (6 군집)	제6군집 (2개)	**Violence**
		Culture			Emotions
		Identity	제2영역 (3 군집)	제7군집 (4개)	**Bourdieu**
		Religion			Ethics
		Education			Reflexivity
		Power			Habitus
		Social Theory		제8군집 (5개)	**Sociology Of Knowledge**
		Politics			Sociology Of Science
		Risk			Epistemology
		Historical Sociology			Actor-Network Theory
		Health			Innovation
		Qualitative Research		제9군집 (4개)	**Ethnography**
		Cultural Sociology			Methodology
		History			Theory
		Discourse			Public Sociology
		Media	제3영역 (4 군집)	제10군집 (2개)	**Economic Sociology**
		Modernity			Markets
		Medical Sociology		제11군집 (7개)	**Globalization**
	제2군집 (5개)	**Gender**			Environmental Sociology
		Class			Neoliberalism
		Embodiment			Consumption
		Social Class			Political Economy
		Family			Sustainability
	제3군집 (4개)	**Inequality**			Climate Change
		Race		제12군집 (2개)	**Political Sociology**
		Ethnicity			Social Movements
	제4군집 (2개)	**Agency**		제13군집 (4개)	**Social Networks**
		Children			Social Capital
	제5군집 (3개)	**Knowledge**			Trust
		Higher Education			Community
		Technology			



〈그림 2〉 논문 기반 사회학 분야의 저자키워드 네트워크 분석(PFNet와 PNNC 세부 군집)

논문을 기반으로 한 사회학 분야는 크게 세 영역으로 구분되며, 우선 전통적인 사회학 분야가 가장 큰 규모로 중심에 형성되어 있다. 전통적인 사회학 분야는 Gender, Culture, Identity, Social Theory, Race 등의 세부 영역으로 연결되어 있다. 두 번째 영역은 Bourdieu, Sociology of Knowledge, Ethnography의 군집을 포함하고 있으며, 사회학 이론과 관련된 군집으로 볼 수 있다. 세 번째 영역은 Economic Sociology, Globalization, Political Sociology, Social Networks로 구성되었다.

4.5 논의

사회학 분야의 연구데이터를 기반으로 한 지적구조와 논문 기반의 지적구조의 비교는 〈표 5〉와 〈표 6〉, 〈그림 1〉과 〈그림 2〉로 설명할 수 있다. 연구데이터를 기반으로 한 사회학 분야의 지적구조와 논문을 기반으로 한 지적구조를 비교한 결과, 세부 군집의 주제 영역이 상당히 다르게 나타났다. 〈표 7〉은 연구데이터를 기반으로 한 사회학 분야와 논문을 기반으로 한 사회학 분야에서 각각 추출된 군집명을 비교하여 정리한 내용이다. 그 내용을 살펴본 결과, 군집명이 일치하거나 유사한 의미를 지니는 경우는 세 가지로 나타났다. Sociology 군집은 연구데이터 기반의 사회학과 논문 기반의 사회학 모두에서 나타났으며, Economics와 Economic Sociology가 서로 근접한 분야로 확인되었으며, Information and Computing Science와 Social Networks 역시 유사한 분야의 군집으로 볼 수 있다.

〈표 7〉 연구데이터와 논문 추출 군집명 비교

번호	연구데이터 군집명	논문 군집명
1	Sociology	Sociology
2	Psychology	Gender
3	Public Health	Inequality
4	Epidemiology	Agency
5	Neurosicence	Knowledge
6	Science Policy	Violence
7	Clinical Research Design	Bourdieu
8	Infectious Diseases	Sociology of Knowledge
9	Ecology	Ethnography
10	Economics	Economic Sociology
11	Communications	Globalization
12	Pediatrics	Political Sociology
13	Information and Computing Science	Social Networks
14	Mathematics	
15	Statistics	

〈표 7〉과 같은 비교를 통해 사회학 분야에서 인용된 연구데이터의 분야와 전통적인 논문에서 다루고 있는 분야가 상당히 차이가 있다는 것을 확인할 수 있다. 특히, 연구데이터를 기반으로 한 사회학 분야의 지적구조를 살펴보면, 의학, 과학, 수학 분야가 큰 두각을 나타난 것을 볼 수 있다. 반면에 논문을 기반으로 한 사회학에서 추출한 군집명은 전통적으로 사회학 분야에서 다루는 영역과 상당히 일치하는 것으로 나타났다. 대학평가기관인 QS에서 발표한 2017년 사회학 세계 대학 순위 상위 5개 대학의 연구자 연구 영역과 비교하였다. 그 결과 군집명 Agency와 Bourdieu를 제외하면 모든 세부 영역의 군집명이 주요 대학의 사회학에서 연구되고 있는 분야로 나타났다.

또한 연구데이터와 논문의 저자키워드를 활용한 동시출현단어 분석 결과로 시각화한 네트워크(〈그림 1〉과 〈그림 2〉) 구성에서도 차이를 찾아볼 수 있다. 우선 공통적으로 두 네트워크에서 Sociology 군집이 가장 중심적인 역할을 하는 것으로 나타났다. 그러나 연구데이터를 기반으로 한 사회학 분야의 네트워크에서 Sociology 군집은 주변의 Medicine, Neuroscience 군집과 연결되어 있다. 또한 Medicine 군집은 Psychology, Public Health 군집과 주제적인 연결을 찾아볼 수 있다. 반면, 논문을 기반으로 한 사회학 분야의 저자키워드 동시출현단어 분석 네트워크에서 Sociology는 Culture, Race, Social Networks, Ethnography 등의 군집과 다양하게 연결되어 있는 것으로 나타났다.

이러한 연구데이터 기반의 사회학 분야와 논문 기반의 사회학 분야 간의 비교 분석을 통해서 데이터의 접근과 재이용을 통해 사회학 분야가 가진 여러 학문 분야와의 연계성과 학제적인 연구경향을 파악하였다.

5. 결 론

본 연구는 인용된 연구데이터와 논문을 기반으로 한 사회학 분야를 비교하고, 데이터 공유와 재사용을 통해 도출된 연구데이터 기반의 사회학 분야가 가지는 특성과 지적구조를 규명하였다.

연구데이터 기반의 사회학 분야의 지적구조는 2영역 15군집으로 나타났다. 세부 영역을 군집별 대표 키워드를 통해 살펴본 결과, 제 1군집 사회학, 제 2군집 심리학, 제 3군집 공중보건학, 제 4군집 역학, 제 5군집 신경과학, 제 6군집 과학기술정책, 제 7군집 임상 연구 설계, 제 8군집 경제학, 제 9군집 전염병, 제 10군집 생태학, 제 11군집 의사소통, 제 12군집 소아과,

제 13군집 정보 시스템 과학, 제 14군집 수학, 그리고 제 15군집 통계학으로 나타났다.

논문을 기반으로 한 사회학 분야는 3영역 17군집으로 구성되어 있다. 세부 영역을 대표 키워드를 통해 살펴본 결과, 제 1군집 사회학, 제 2군집 젠더, 제 3군집 불평등, 제 4군집 에이전시, 제 5군집 지식, 제 6군집 폭력, 제 7군집 부르디외, 제 8군집 지식사회학, 제 9군집 민속학, 제 10군집 경제사회학, 제 11군집 세계화, 제 12군집 정치사회학, 그리고 제 13군집 소셜네트워크로 나타났다.

연구데이터를 기반으로 한 사회학 분야의 특성과 지적구조를 논문을 기반으로 한 사회학 분야와 종합적으로 비교하여 분석한 내용은 다음과 같다. 첫째, DCI 전체 데이터 중 약 88%가 인용되지 않았다는 것과 비교하였을 때, 전체 데이터의 약 86%가 1회 이상 인용이 된 사회학 분야의 데이터는 데이터인용이 활발히 이루어지고 있는 분야이다. 둘째, 연구데이터를 기반으로 한 사회학 분야는 전통적인 사회학에서는 세부적으로 다뤄지지 않았던 의학 분야가 활발히 연구 중인 것으로 나타났다. 특히 의학 분야 중에서도 공중보건과 심리학이 중심이 되고 있는 것으로 나타났다.

본 연구는 데이터인용 색인 데이터베이스인 DCI의 데이터를 활용하여 사회학 분야의 데이터 공유와 재이용을 통해 구축된 새로운 지적구조를 규명하여 새로운 학문적 시각을 제시하였다는 의의를 가진다. 또한, 오픈사이언스와 오픈데이터라는 새로운 연구기반 환경에서 데이터의 공유와 재이용이 학문 분야에 미치는 영향을 조명한 새로운 시도로, 연구데이터를 기반으로 한 새로운 연구의 방향성을 모색하기 위한 기반이 될 수 있을 것이다.

참 고 문 헌

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

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. PloS One, 9(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Dalerba, P., Sahoo, D., Paik, S., Guo, X., Yothers, G., Song, N., .… Hisamori, S. (2016). CDX2 as a prognostic biomarker in stage II and stage III colon cancer. New England Journal of Medicine, 374(3), 211-222. https://doi.org/10.1056/nejmoa1506597

Diekema, A. R., Wesolek, A., & Walters, C. D. (2014). The NSF/NIH effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. The Journal of Academic Librarianship, 40(3), 322-331. https://doi.org/10.1016/j.acalib.2014.04.010

European Commission (2015). Study on open science: Impact, implications and policy options. Retrieved from https://ec.europa.eu/research/innovation-union/pdf/expert-groups/rise/study_on_open_science-impact_implications_and_policy_options-salmi_072015.pdf

Lo, B. (2015). Sharing clinical trial data: Maximizing benefits, minimizing risk. JAMA, 313(8), 793-794. https://doi.org/10.1001/jama.2015.292

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship and Scholarly Communication, 1(1), 1-16. https://doi.org/10.7710/2162-3309.1035

National Cancer Institute (2006). Data sharing policy. Retrieved from https://ctep.cancer.gov/protocolDevelopment/docs/data_sharing_policy.pdf

National Institutes of Health (2003). NIH data sharing policy and implementation guideline. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Onyancha, O. B. (2016). Open research data in Sub-Saharan Africa: A bibliometric study using the Data Citation Index. Publishing Research Quarterly, 32(3), 227-246. https://doi.org/10.1007/s12109-016-9463-6

Parr, C. S., & Cummings, M. (2005). Data sharing in ecology and evolution. Trends in Ecology & Evolution, 20(7), 362-363. https://doi.org/10.1016/j.tree.2005.04.023

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227. https://doi.org/10.1126/science.1213847

QS World University Rankings (2017). Retrieved from https://www.topuniversities.com/university-rankings/university-subject-rankings/2017/sociology

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2015). Analyzing data citation practices using the Data Citation Index. Journal of the Association for Information Science and Technology, 67(12), 2964-2975. https://doi.org/10.1002/asi.23529

Ross, J. S., Tse, T., Zarin, D. A., Xu, H., Zhou, L., & Krumholz, H. M. (2012). Publication of NIH funded trials registered in ClinicalTrials.gov: Cross sectional analysis. BMJ, 344, d7292. https://doi.org/10.1136/bmj.d7292

Simberloff, D., Barish, B. C., Droegemeier, K. K., Etter, D., Fedoroff, N., Ford, K., … White Jr, J. A. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Virginia: National Science Foundation.

Spires-Jones, T. L., Poirazi, P., & Grubb, M. S. (2016). Opening up: Open access publishing,

data sharing, and how they can influence your neuroscience career. European Journal of Neuroscience, 43(11), 1413-1419. https://doi.org/10.1111/ejn.13234

Stodden, V., Leisch, F., & Peng, R. D. (Eds.). (2014). Implementing reproducible research. New York: CRC Press.

이재윤. COOC ver 0.4 프로그램 [cited 2017.02.06.]

이재윤. WNET ver 0.4.1 프로그램 [cited 2017.02.06.]

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

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