연구 데이터의 출판과 인용

2017. 02.

한국과학기술정보연구원

〈 목 차 〉

Ⅰ 데이터 출판 개요

1 데이터 출판의 정의와 대상

(1) 정의

- 데이터 출판이란 다른 사람들이 (재)활용할 수 있도록 데이터를 일반에 공개하는 행위로서 데이터(또는 데이터 셋)를 일반인들이 자신들이 원하는 목적으로 이용할 수 있도록 하는 것을 뜻함[1)]

(2) 배경

- 오픈 사이언스 운동[2)]에 따라 연구 과정에서 생성된 관련 데이터의 공개와 재활용에 대한 요구가 증가하고 있음
- 정부 부처, 연구기관, 연구비 관리기관 등이 연구성과물의 공개를 요구하는 경향이 증가하고 있음
- 학회, 대형출판사 등은 논문 게재시 관련 데이터의 제출을 함께 요청하는 경우가 증가하고 있으며 각종 학술단체에서는 연구자들에게 연구 관련 데이터를 공개하거나, 인용과 출판을 권고하고 있음

(3) 목표

- 데이터 출판의 목표는 과학적 연구가 재생산되고 데이터가 재이용되도록 하기 위한 것으로서, 연구 논문의 근거가 되는 데이터를 널리 출판한다는 것은 논문의 부정행위뿐만 아니라 정직한 오류를 공개할 수가 있으며, 출판된 학술 논문의 근거가 되는 데이터의 투명성으로 인해 과학적 재생산성을 향상시킬 수가 있음
- 데이터를 적절하게 재이용하게 되면 연구비용을 낮추고 연구를 촉진시킬 수 있는데, 데이터의 문서화, 출판과 저장에는 시간과 비용이 소요되나 데이터를 반복해서 생산하는 비용에 비하면 훨씬 저렴하며 같은 비용을 투자할 경우 논문 생산에 비해 데이터를 정제, 저장, 출판하는 편이 훨씬 학문 발전에의 파급 효과가 크다는 것이 보고되고 있음
- 또한 어떤 데이터들은 재생산에 많은 비용이 소요될 뿐만 아니라 시간 경과에 따라 측정한 기후 데이터나 특정한 시점에서 관찰한 우주 현상 데이터는 돈으로 대체할 수 없는 귀한 가치를 가지고 있음

(4) 대상

- 실험, 측정, 관측, 계산 등 연구 개발 과정에서 생성, 가공, 재생산된 모든 연구 데이터를 출판 대상으로 함

---

1) https://en.wikipedia.org/wiki/Data_publishing (2016. 2.)
2) https://en.wikipedia.org/wiki/Open_science (2016. 2.)

(5) 이익과 효과

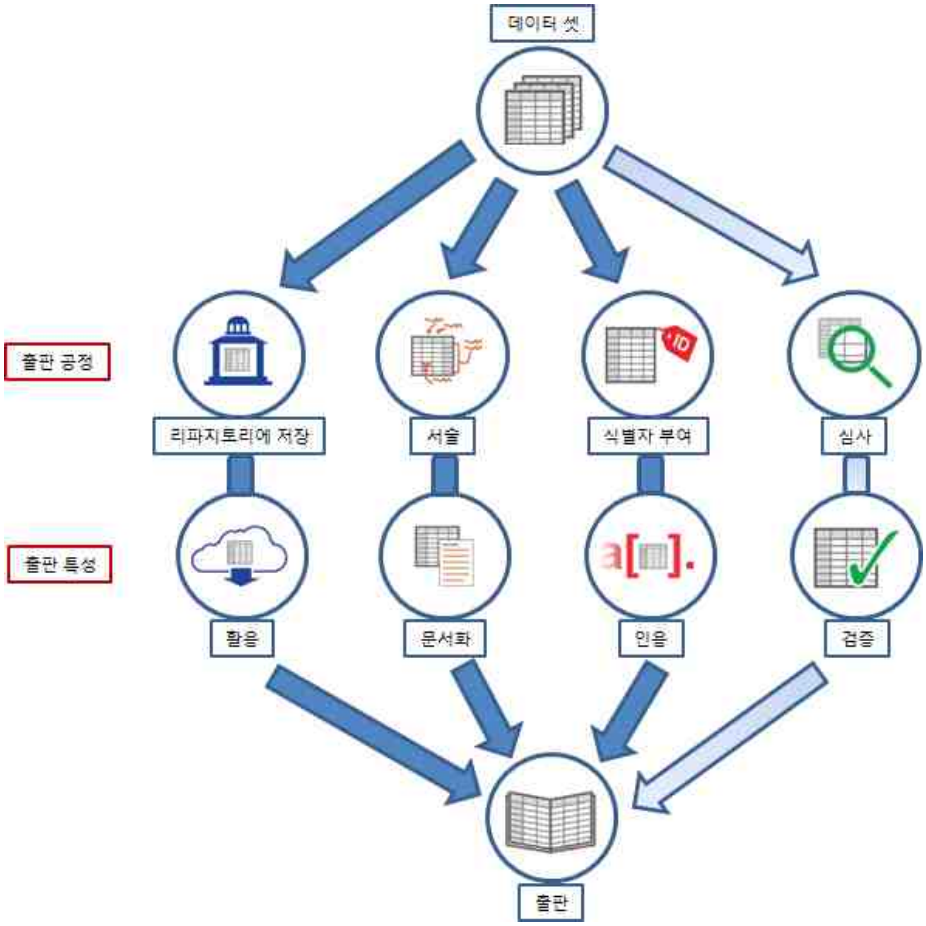
- 자신의 데이터를 다른 연구자가 검색해서 이용할 수 있음
- 다른 연구자가 자신의 데이터를 어떻게 이용했는지 추적할 수 있음
- 정부 또는 연구비 지원기관에서 요청하는 데이터관리계획(DMP)에 데이터 출판 내용을 작성하여 제출함으로써 데이터관리계획서 제출 의무를 수행할 수 있음
- 자신의 데이터에 DOI를 부여받을 수 있어서 데이터의 접근과 인용을 보다 용이하게 할 수 있음
- 연구자가 데이터 출판 과정에 직접 참여함으로써 데이터 가공, 저장, 인용 등에 관한 정보나 데이터 출판과 관련된 지원과 인센티브를 받을 수 있음
- 데이터의 출판과 인용을 통해 연구자 간에 데이터 재활용이 촉진되고 데이터가 학술 논문과 동일한 성과물로서 인정을 받게 되며[3] 국가적으로는 데이터 인프라 구축을 통해 과학기술 혁신활동에 기여할 수 있음

3) Callaghan, S. etal, (2012). "Making data a first class scientific output: Data citation and publication by NERCs environmental data centres". International Journal of Digital Curation 7 (1): 107-113. doi:10.2218/ijdc.v7i1.218

II 데이터 출판 요소

1 데이터 출판의 특성

- 학술 단체에서 일반적으로 동의하고 있는 데이터 출판의 3가지 중요한 특성은 <그림 1>과 같이 표현할 수 있으며 그 내용은 다음과 같음[4)]



<그림 1> 데이터 출판을 위한 리파지토리 저장, 문서화, 식별자 부여와 심사

① 활용성 : 출판된 데이터는 현재와 미래에 일반인들이 활용할 수 있어야 하며, 데이터의 접근에 비용이 소요되거나 동의서를 작성해야 할 수도 있지만 데이터에 대한 권리를 저자들에게 한정해서는 안됨
② 문서화 : 출판된 데이터는 적어도 동일 분야의 연구자들이 재생산 또는 재활용할 수 있을 정도로 적절하게 문서화되어야 함
③ 인용 : 서적이나 저널 논문과 같이 데이터 출판도 공식적으로 인용되어야 하며,

4) F1000Research 2014, 3:94 Last updated: 09 MAR 2016, doi:10.12688/f1000research.3979.3

데이터 인용은 학술적 출판물과의 통합을 유지할 뿐만 아니라 연구자로 하여금 데이터를 출판하도록 동기를 부여하며 보상을 줄 수가 있음

④ 검증 : 출판사들은 데이터를 검증하기 위해 심사 과정을 거치는 경우가 있으며 일반적으로는 공개된 질문들에 의해 출판된 데이터가 검증됨

2 활용성

- 데이터를 출판한다는 것은 데이터를 일반인이 활용할 수 있도록 한다는 의미이며, 지금 활용하기 위해서는 현재 데이터에 접근할 수 있어야 하고 미래의 활용을 위해서는 장기 저장이나 포맷 변환과 같은 데이터 보존 작업이 필요함
- 인쇄 출판물처럼 출판된 데이터가 유료이고 법적으로 보호되고 있으며 이용 동의서가 필요하다면 출판 데이터의 활용성은 크게 제한받게 되며, 데이터에 대한 접근이 제한되면 투명하고 객관적인 데이터 판단 기준을 적용할 수 없음
- 가장 일반적으로 데이터 접근을 제한하는 이유는 개인정보 보호를 위한 것이며 특히 의료정보는 이런 의미에서 공개를 제한하고 있음
- 한편 데이터를 출판하기 위해서는 신뢰성이 있는 리파지토리에 데이터를 저장해야 하며, 많은 기관들이 신뢰성 있는 리파지토리를 평가하기 위한 체크리스트[5]를 발표하고 있음
- 실제로는 리파지토리를 관리하는 관리기관이 리파지토리의 신뢰성을 좌우하게 되며, 정부기관이나 대학들이 운영하는 리파지토리가 신뢰성이 있다고 생각됨

3 문서화

- 데이터가 활용되어 연구가 재생산되기 위해서는 메타데이터 작성과 같이 데이터에 대한 문서화가 이루어져 데이터의 내용을 이해할 수 있어야 함
- 데이터의 문서화에는 시간과 노력이 필요하고 연구실 내의 데이터가 연구실 밖에서 활용될 수 있도록 하는 것이며 이 노력에는 보상이 따라야 데이터 출판이 촉진될 수 있음
- 데이터의 문서화는 관련 학술문헌과 데이터를 서로 연결할 수 있는데 하나의 데이터가 많은 문헌과 관련이 있을 수도 있고 하나의 문헌이 많은 데이터를 설명할 수도 있으며, 데이터는 문헌과 관련하여 다음과 같이 3가지로 문서화될 수 있다.

(1) 연구 논문을 보조하는 데이터

- 데이터 출판의 가장 친근한 형태가 전통적인 저널 문헌에 보조 자료로서 동반되

5) Center for Research Libraries (U.S.) and OCLC. Trustworthy repositories audit & certification (TRAC) criteria and checklist. Center for Research Libraries ; OCLC Online Computer Library Center, Inc Chicago: Dublin, Ohio. 2007.

는 데이터의 출판임
- 이러한 데이터는 저널이 보조 자료로서 보유하거나 제3의 리파지토리에 저장이 되는데 최근 들어 리파지토리가 장기 보존과 데이터 접근에 적절하다고 판단되어 저널이 보조 자료로서 데이터를 출판하는 경향은 줄어들고 있음
- 예를 들어 The Journal of Neuroscience는 2010년에 보조 자료로서 데이터 출판을 중지했으며 분야별 리파지토리가 데이터를 유통시키는데 훨씬 적합하다고 공지하고 있음
- 심사 과정이 있는 데이터의 출판은 Dryad[6] 리파지토리에 저장되고 있는데 Dryad는 데이터를 활용하고 인용할 수 있도록 하지만 문헌 출판자는 데이터에 대해 과학적인 검증 절차를 거쳐야 하며, Research Compendia[7]는 출판된 논문과 그 근거가 되는 데이터나 코드를 모두 저장하고 있음
- 최근에는 논문과 관련된 데이터 리파지토리가 아니더라도 많은 출판사들이 논문의 근거가 되는 데이터 출판을 권유하고 있으며, 이러한 종류의 데이터 출판은 연구분석 활동의 재생산을 지원하기 위한 것이며 반드시 데이터의 재이용을 위한 것은 아님
- 예를 들면 PLOS[8] 데이터 정책은 논문의 발견의 재생산에 필요한 데이터만 출판하도록 하고 있으며 수집된 모든 데이터의 출판이나 관련 없는 목적의 데이터 재이용을 위한 문서화는 하지 않는다고 규정하고 있음

(2) 논문 주제로서의 데이터
- 데이터 논문은 데이터의 상세한 수집 방법에 대해서 서술하며 분석이나 결론이 없음.
- 데이터 논문은 F1000Research[9], Internet Archaeology[10], GigaScience[11]와 같은 저널에 새로운 논문 형태로서 활발히 발표되고 있으며 전문적인 데이터 저널로서는 Earth System Science Data[12], Geoscience Data Journal[13], Nature Publishing Group의 Scientific Data[14], Ubiquity Press의 metajournals 트리오가 있음
- 데이터 논문의 강점은 풍부한 문서화에 있는데 특히 크기가 작고 독특한 특징이 있는 작은 용량의 연구 데이터(롱테일 연구 데이터)에 매우 유용함
- 데이터 논문의 길이와 구조는 저널마다 다르나 일반적으로 모든 저널은 초록과 데이터의 수집 방법, 데이터에 대해 서술하도록 하며, 일부는 데이터의 잠재적 용도를 제시하도록 한다든가(예를 들면 Internet Archaeology, Open Health Data), 분야 고유의 섹션을 추가하는 일도 있음(예를 들면 Internet Archaeology와

6) http://datadryad.org/, (2016. 9.)
7) http://researchcompendia.org/, (2016. 9.)
8) https://www.plos.org/, (2016. 9.)
9) http://f1000research.com/, (2016. 9.)
10) http://www.internetarchaeology.org/index.htm, (2016. 9.)
11) http://gigascience.biomedcentral.com/, (2016. 9.)
12) http://www.earth-system-science-data.net/, (2016. 9.)
13) http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060, (2016. 9.)
14) http://www.nature.com/sdata/, (2016. 9.)

Journal of Open Archaeology Data는 시간과 지리 정보 섹션을 포함함).

- 데이터 저널은 일반적으로 데이터에 대한 서술을 출판하는 것을 제한하는 일이 있으나 신뢰성이 있는 리파지토리는 데이터 자체를 출판함
- 예를 들면 Scientific Data와 Geoscience Data Journal은 저자들에게 신뢰성 있는 리파지토리 목록을 제시하며 GigaScience는 GigaDB라는 통합 리파지토리에 데이터를 저장하도록 하고 The International Journal of Robotics Research[15]는 저자들로 하여금 자신들의 웹사이트에 데이터를 저장하도록 함

(3) 독립적인 문서화

- 많은 리파지토리들은 문헌의 참조 없이도 데이터를 재이용하거나 연구의 재생산이 가능하도록 데이터를 문서화 하는데, 대학이나 정부 또는 민간기업에서 만든 리파지토리가 연구 커뮤니티를 중심으로 하여 논문과 연계되거나 그렇지 않은 형태로 다양한 프로세스를 통해 데이터를 출판하고 있음
- 기관 리파지토리는 연구 커뮤니티에서 생성된 다양한 종류의 데이터를 보존하고 출판하는데, 예를 들면 캘리포니아 대학 연구자들은 데이터를 Merritt[16]에 저장하고 퍼듀 대학은 Purdue Research Repository(PURR)[17]를 이용하고 있으며, 국가 차원에서는 네덜란드가 연구자들로부터 광범위한 데이터를 Dutch Data Archiving and Networked Services(DANS)[18]를 통해 받고 있음
- Figshare[19]와 Zenodo[20]는 모든 연구자로부터 모든 분야에 걸쳐 데이터를 받아 출판하고 있으며 이러한 광범위한 주제를 담당하는 데이터 출판자들은 특화된 전문 분야의 리파지토리에는 맞지 않는 롱테일 데이터나 불규칙한 데이터들을 취급할 수가 있으나, 취급 분야가 광범위하기 때문에 분야 특유의 전문적인 문서 작성이나 검증에는 한계가 있으며 따라서 최소한의 메타데이터 문서화를 요구하고 있음
- 한편 관심 주제의 연구 커뮤니티를 지원하는 특화된 데이터 출판자들이 있는데 이 출판자들은 해당 관심 분야의 모든 정보들을 출판하고 있음
- 예를 들면 고고학 기록을 다루는 the Digital Archaeological Record(tDAR)[21]이 있고, 생명과학 중에서는 좁은 범위의 모델 생물 데이터베이스 그룹인 WormBase[22]과 FlyBase[23], 그리고 유전자 발현 데이터를 수집하는 Gene Expression Omnibus(GEO)[24]이 있으며, 지진 데이터를 수집하는 SeismicPortal[25]이 있음
- 이러한 특화된 분야의 데이터를 이용할 수 있도록 하기 위해서는 기술적 검증과

15) http://www.ijrr.org/, (2016. 9.)
16) https://merritt.cdlib.org/, (2016. 9.)
17) https://purr.purdue.edu/, (2016. 9.)
18) https://dans.knaw.nl/en, (2016. 9.)
19) https://figshare.com/, (2016. 9.)
20) https://zenodo.org/, (2016. 9.)
21) http://www.tdar.org/, (2016. 9.)
22) http://www.wormbase.org/#012-34-5, (2016. 9.)
23) http://flybase.org/, (2016. 9.)
24) http://www.ncbi.nlm.nih.gov/geo/, (2016. 9.)
25) http://www.seismicportal.eu/, (2016. 9.)

특화된 메타데이터의 개발이 필요한데 예를 들면 GEO 데이터는 마이크로어레이 실험에 대한 최소한의 정보(Minimum Information About a Microarray Experiment, MIAME) 문서화 지침에 의해 메타데이터를 작성하고 있음

- 데이터 출판자가 특수한 과학적 장비나 시설을 보유하는 경우가 있는데 예를 들어 One Degree Imager Portal, Pipeline, and Archive (ODI-PPA)[26]나 강입자 충돌기의 생성 데이터를 취급하는 거대한 기반시설인 Worldwide LHC Computing Grid[27]가 있으며, 이 경우에는 다른 출판자들과는 달리 장비에서 생성되는 데이터의 실시간 접근을 강조하고 있음
- 분야별 리파지토리에 있는 데이터들은 해당 분야의 과학자들이 대부분 알고 있기 때문에 쉽게 발견할 수 있으며 대부분이 표준화되어 있어서 재이용하기가 쉬움
- 다만 단일 연구과제에서 생성된 데이터들이 많은 리파지토리에 나뉘어 유통되는 단점이 있는 반면에(예를 들면 유전자 발현정보와 시퀀스 데이터가 각각 다른 리파지토리에 저장) 기관 리파지토리나 광범위한 주제의 일반 리파지토리는 연구과제 전체를 출판할 수가 있음

4 인용

- 데이터 인용은 데이터 출판의 한 요소이지만 의견의 일치를 이루기가 가장 어려운 부분으로서, 2014년 초에 Future Of Research Communication and E-Scholarship(FORCE11) 연합 조직이 Joint Declaration of Data Citation Principles[28]이란 데이터 인용 선언문을 발표하였음
- 여기에서 “학술지에 있어서 데이터의 인용은 학술 출판물의 인용과 동일한 수준의 중요성을 가져야 한다”고 데이터 인용의 중요성을 강조했는데 이것은 데이터가 공식적으로 참고 문헌 리스트에서 인용되어야 한다는 것을 뜻함
- 그러나 실제로는 출판사들이 참고 문헌에 데이터 인용을 허락하지 않으며, 설사 허락하여도 저자들이 공식적인 인용이 아니고 본문 안에서 데이터를 참고할 뿐임
- 최근에 명확한 인용 지침을 제공하는 데이터 출판자들이 증가하고 있는데, Dryad[29], Figshare[30], Zenodo[31]의 데이터 셋 랜딩 페이지에는 데이터의 인용 형식이 표시되어 있음
- 많은 데이터 출판자들은 DOI와 같은 영구식별자를 데이터에 부여함으로써 데이터를 인용하기 쉽도록 지원하고 있으며, DOI를 통해 인용된 데이터의 상세 내용과 소재지를 파악할 수 있음

26) https://portal.odi.iu.edu/index/front, (2016. 9.)
27) http://wlcg.web.cern.ch/, (2016. 9.)
28) https://www.force11.org/group/joint-declaration-data-citation-principles-final, (2016. 9.)
29) http://datadryad.org/ (2017. 2.)
30) https://figshare.com/ (2017. 2.)
31) https://zenodo.org/ (2017. 2.)

- 그러나 DOI가 인용에 반드시 필요한 것은 아니며 데이터의 소재지가 변경되었을 때 DOI도 같이 갱신되지 않으면 의미가 없음

(1) 단순 인용

- 현재 데이터 인용은 문헌 인용과 유사한 최소한의 5개 요소(생성자, 제목, 연도, 출판자, 식별자)를 사용하여 인용하고 있으며, 이 형식은 Committee on Data for Science and Technology(CODATA)[32]에서 추천한 것으로서 DataCite[33]의 DOI를 취득하거나 Thomson-Reuters Data Citation Index[34]에 실리기 위해 필요한 정보임

(2) 심층 인용

- 대규모 데이터 셋의 서브 셋 데이터를 사용하여 수행한 분석을 재현하기 위해서는 정확하게 어떤 서브 셋 데이터가 사용되었는지를 알 필요가 있는데, 데이터 셋은 구조적으로 광범위하게 변화하기 때문에 세브 셋 데이터를 서술할 수 있는 일반적인 해법이 없으며 가장 많이 통용되는 것으로는 참고 문헌 목록에 모든 데이터 셋을 인용하고 서브 셋 데이터는 논문의 본문 안에서 서술하는 것임
- 세브 셋 데이터는 내부 구조(예를 들면 시간 또는 공간 범위, 변수 목록, 내부 식별자 등)에 맞는 형식을 사용하여 인용하기를 권장하고 있음

(3) 역동적 데이터 셋

- 과거에 인쇄 공정은 논문을 하나의 기록물로서 하나의 버전으로 고정시켰지만 웹기반 출판이나 프리프린트 서버(예를 들면 arXiv.org[35])에서의 상황은 매우 복잡하며 특히 데이터 셋이 역동적으로 변경되는 경향이 많이 있음
- 역동적 데이터 셋에는 두 가지 종류가 있는데 하나는 기존의 데이터에 신규 데이터가 추가되는 성장 데이터 셋이고 다른 하나는 데이터가 추가, 삭제, 변경되는 수정 데이터 셋이 있음
- 출판된 데이터 셋에 데이터를 새로 추가하는 것은 매우 가치 있는 일이며 역동적 데이터 셋을 재해석하고자 하는 연구자들은 특정한 버전의 데이터에 접근하고 싶어 함
- 따라서 성장 데이터 셋은 접근 날짜나 인용의 유효 기간 범위와 함께 인용해야 하며, 수정 데이터 셋은 개정된 버전을 모두 보관하면서 인용할 수 있는 버전 번호와 함께 주기적으로 새로운 버전을 출판함으로써 연구자는 새로운 버전의 데이터와 함께 과거 버전의 데이터도 모두 이용할 수 있도록 해야 함

32) http://www.codata.org/, (2016. 9.)
33) https://www.datacite.org/, (2016. 9.)
34) http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/data-citation-index.html, (2016. 9.)
35) http://arxiv.org/, (2016. 9.)

5 데이터 심사

- 데이터의 심사에는 기술적 심사와 과학적 심사가 있는데, 기술적 심사란 데이터 셋과 그 서술 내용이 완전하고 서로 잘 맞는지를 확인하는 것이며 여기에는 해당 분야의 전문가가 필요하지 않기 때문에 많은 리파지토리들이 데이터 셋에 대해 어느 정도의 기술적 심사를 실시하고 있음
- 한편 과학적 심사란 데이터의 수집 방법, 데이터의 전반적인 타당성, 재이용 가치와 가능성 등을 평가하며 해당 분야의 전문가를 필요로 하기 때문에 심사는 까다로우며, 데이터가 데이터 논문과 함께 출판될 경우 심사 과정은 기술적 심사를 위한 리파지토리와 과학적 심사를 위한 데이터 저널의 두 가지로 나뉘게 됨

(1) 데이터 논문의 심사

- 데이터 논문 저널은 하나의 과정을 통해 논문과 데이터 셋의 과학적 심사를 진행하며, GigaScience는 데이터의 기술적 심사를 별도의 데이터 심사자에게 의뢰하고 있음
- 많은 데이터 저널이 심사자에게 가이드라인을 제공하는데 독창성이나 잠재적 효과를 고려하거나 데이터 셋이 과학적으로 확실한 것인지를 검증하며, 가이드라인이 유사하다고 해도 심사 과정은 매우 다양함
- 예를 들어 Biodiversity Journal과 Scientific Data를 비교하면 두 저널은 심사자의 가이드라인을 “데이터의 품질”, “서술 내용의 품질”, “원고와 데이터 사이의 일관성”의 3가지 부분으로 나누고 있음
- Scientific Data는 편집자가 익명의 심사자를 지정하는데 비해 Biodiversity Journal은 심사자의 익명이 선택 사항이며 여러 유형의 심사자를 지정하는 유연하고 오픈된 심사 방식을 따르고 있음. 여기에서 편집자는 심사 보고서를 제출해야 하는 2-3명의 추천 심사자와 논문을 읽고 단지 자신의 의견만을 내는 몇 명의 패널 심사자를 지정하며, 저자는 심사 과정 중에 일반인의 심사 의견을 구하기 위해 자신의 논문을 공개할 수도 있음

(2) 독립적인 데이터의 검증

- 데이터 저널이 데이터 심사를 위해 사용한 대부분의 모델이 문헌의 심사 과정을 따르는데 비해 독립적인 데이터 검증 사례와 제안은 매우 다양함
- 예를 들면 데이터 저널의 가이드라인과 유사한 독립적인 데이터 심사 가이드라인이 2011년에 발표되었는데 여기에서 NASA의 Distributed Active Archive Centers(DAACs)[36]는 특정 분야 전문가로 구성된 이용자 워킹 그룹을 운영하고 있으며, NSIDC는 기술적 심사에 해당하는 서비스 수준의 데이터 출판에 필요한 내부 심사와 과학적 품질을 심사하는 외부 심사를 병용하고 있음. 또한 Planetary Data System(PDS)[37]은 리파지토리의 대표자와 데이터 셋의 생산자, 그

36) http://www.nasa.gov/, (2016. 9.)

리고 심사자들이 모인 회의를 통해서 데이터 셋을 심사하고 있음
- 출판 전의 데이터 심사는 출판 후에 데이터 이용자들의 피드백에 의해 보완될 수 있는데, 데이터를 재사용하는 것은 일종의 검증과 같아서 데이터가 단순히 좋다기보다는 어떤 특정한 목적에 맞다고 검증할 수 있기 때문임
- DANS 리파지토리는 데이터를 사용하고 있는 연구자들에게 피드백을 요청하는데 그 내용은 데이터 품질, 문서화의 품질, 데이터 셋의 구조 등과 같은 6개의 각 평가 지표에 대해 1에서 5까지 점수를 매기도록 하고 있음. 연구자들은 데이터의 생성 과정과 한계를 알기 때문에 이러한 검증과정을 신뢰하며 연구자들이 데이터를 이해하게 되면 출판 전 또는 후의 검증 과정은 데이터 심사와 동일한 수준으로 간주할 수 있음
- 고고학 분야에서 Open Context와 the Digital Archaeological Record(tDAR)의 두 가지 사례를 통해 데이터 검증의 여러 가지 방법을 알 수 있음.
- Open Context는 심사를 통한 데이터 검증 과정을 갖추고 있는데, 개개의 Open Context 데이터 셋은 품질 자체뿐만 아니라 검증의 완전성을 평가 지표로 하여 1에서 5까지의 점수를 부여함. 1점은 보증할 수 없다는 뜻이며, 3점은 기술적 심사를 통과했고, 5점은 외부의 심사 과정도 통과했다는 뜻임
- Open Context가 데이터 제출과 재이용에 중점을 둔 소규모의 출판자라면 tDAR은 고고학 데이터를 미래에 활용할 수 있도록 수집하고 보존하는데 비중을 둔 대규모 리파지토리라고 볼 수 있는데, tDAR은 최소한의 필수 작성 항목의 메타데이터와 함께 기술적 심사와 간소화된 데이터 저장만으로 리파지토리를 운영하고 있음
- 그러나 tDAR은 고품질의 데이터 출판 플랫폼으로서 많은 정보를 제공하는 기여자들과 디지털 큐레이터들이 함께 리파지토리를 운영하고 있으며, 데이터 논문 저널인 Internet Archaeology와 Journal of Open Archaeological Data은 모두 tDAR과 Open Context[38]를 심사가 끝난 데이터의 저장 리파지토리로서 사용할 것을 권유하고 있음.
- 이와 같이 데이터 검증은 분야나 데이터 유형뿐만 아니라 기관의 목표와 함께 일하는 연구자를 포함한 외적 요인에도 의존하고 있다고 할 수 있음

37) https://pds.jpl.nasa.gov/, (2016. 9.)
38) https://opencontext.org/, (2016. 9.)

III 데이터 출판 프로세스

- 데이터를 출판한다는 것은 종래의 저널을 출판하는 것과 유사하다고 할 수 있는데, 저자들은 다음과 같이 단계별로 데이터의 출판 공정을 따라 데이터 셋들을 준비, 제출, 심사, 수정, 문서화 함으로써 데이터 출판에 기여하게 됨[39)]

1 출판 계획 수립

- 정제되고 일관성이 있는 데이터 셋이 분석하고 출판하기에 용이하기 때문에 아래와 같이 오류를 줄일 수 있는 전략을 개발해야 하며, 연구비 지원기관에서 요구하는 DMP(데이터관리계획)도 준비해야 함

(1) 사전 심사
- 모든 데이터 셋은 출판되기 전에 해당 분야의 전문가에 의해 품질에 대한 심사를 받게 되는데, 데이터 셋은 논문처럼 창의성이 없다고 해서 거절하지는 않으며 그 이유는 우수한 품질의 데이터 셋은 미래에 예상하지 못한 가치를 지닐 수도 있고 전혀 다른 분야에 활용되거나 통계적으로 활용될 수도 있기 때문임
- 데이터 셋의 심사 기준으로는 아래와 같은 것들이 있음
  ① 데이터 셋의 획득 방법과 품질의 건전성
  ② 데이터 셋의 문서화에 대한 품질
  ③ 보다 넓은 분야에의 재이용에 대한 적합성

(2) 데이터 준비
- 데이터 출판에는 보다 강화된 데이터 가공과 검증 작업이 요구되는데 아래와 같은 사항들을 체크하면 이러한 작업을 보다 효율적으로 수행할 수 있음

1) 우수한 데이터베이스 기획
- 연구 과제의 착수 시점에서 데이터베이스를 잘 기획하면 생성되는 데이터를 보다 쉽게 출판할 수가 있음
- 데이터의 품질을 유지하기 위해서는 표준화와 일관성이 중요한데 예를 들면 수치 데이터 항목에는 수치 데이터만 입력되어야 하며 수치 정보에 추가적으로 주석 또는 설명이 필요할 경우에는 다른 항목을 만들어서 기입해야 함.
- 데이터 수집의 모든 단계에서 에러 체크 또는 품질 검증 절차가 있다면 데이터는 더욱 신속하게 출판될 수 있고 출판된 데이터는 더욱 가치 있고 많은 이용자들이 쉽게 이용할 수 있게 될 것임

39) http://opencontext.org/about/publishing, (2016. 10.)

2) 정제와 편집
- 데이터 출판은 출판물의 형태이나 저널 논문이나 서적과는 그 특성이 다름. 대부분의 데이터 셋은 원시 데이터 형태로 생성되므로 이를 활용하기 위해서는 1차, 2차의 가공이 필요하고 데이터를 설명하는 용어도 일관성이 있어야 하며, 식별자나 개요 설명에서도 단어, 기호 등에 오류가 없어야 함

3) 개요 설명
- 데이터 셋의 모든 작성 항목은 한두 문장이 되더라도 이해하기 쉽도록 서술적으로 작성해야 함. 특히 작성 항목에 일반적으로 많이 사용되지 않는 용어나 값들을 기입하였을 경우에는 반드시 이들을 이해할 수 있도록 설명해야 함.

4) 관계 구조
- 데이터들이 복잡한 구조를 가진 관계형 데이터베이스로 관리될 때에는 이러한 구조들을 잘 설명해 줌으로써 편집자들이 데이터를 용이하게 추출할 수 있도록 함

5) 소재지와 콘텐츠
- 각각의 데이터 콘텐츠와 그 소재지, 관련 인물과 출판되는 데이터 파일에 대해 해당 URL 주소나 식별자를 설명하는 항목과 그 내용이 상세하게 기술되어야 편집자에게 도움이 될 수 있음

6) 영상과 미디어 파일
- 영상, 미디어 파일 등은 데이터의 중요한 요소이며 이러한 데이터 셋에 대한 메타 데이터는 충실하게 작성되고 데이터 관련 사항(생성자, 기여자, 소재지, 데이터 콘텐츠, 식별자, 관련 URL, 참고문서 등)이 명확해야 함.

7) 연구 개요
- 각각의 개별 연구과제들은 연구 목적과 내용, 연구 방법, 예상되는 생성 데이터와 잠재적 이용자 등을 설명하고 있는 연구 개요를 상세하게 작성해야 함

8) 기여자
- 데이터 인용을 위해 메타 데이터에는 데이터 생성과 가공에 기여한 기여자를 한 명 이상 반드시 기재해야 하며 어떤 경우에는 데이터의 관측자와 분석자도 별도의 항목으로 기재할 수 있음.
- 기여자의 이름은 약어가 아닌 전체 이름을 철자의 오류 없이 정확하게 기입해야 하며 표준이 있을 경우에는 이에 따르고, 이름은 데이터에 1차 책임이 있는 사람을 우선적으로 기입하며 소속기관을 함께 적음

(3) 소재지 정보와 현장 안전
- 프로젝트에 따라서는 지리정보를 제공해야 하는 경우가 있는데 이 지리정보에는 소재지를 알아 볼 수 있도록 적절한 소재지 정보가 포함되며 대개 현장의 장소를 의미함. 그러나 소재지를 밝힐 경우에는 그 현장의 안전 문제를 고려해야 하며 소재지를 밝힘으로써 그 현장의 안전 문제에 위협이 된다면 대외적으로는 정확도가 떨어지는 불확실한 정보를 제공하며 이용자에게 이러한 사실을 알리고 정확한 소재지 정보를 문의할 수 있는 연락처를 명기해 놓는 것이 좋음

2 출판 신청

- 데이터 출판을 위해서는 출판 계획서와 함께 출판 신청을 해야 하며, 출판계획서에는 생성되는 데이터 셋의 주제 분야, 연구 주제, 데이터 셋의 크기와 복잡도, 예상되는 이용자(일반인과 전문가), 데이터 민감도(개인정보, 기밀 등), 저작권이 허용하는 한도 내에서 데이터 출판에 대한 동의서, 출판 비용 등의 내용이 포함되어야 함

(1) 저작권과 라이선싱

- 데이터 생산자들은 미래의 연구를 위해 자신들의 데이터가 법적으로 사용될 수 있도록 해야 하며, 데이터의 법적인 재사용을 가능하도록 하기 위해서는 모든 데이터는 저작권의 간섭 없이 일반에게 공개되거나 크리에이티브 컴먼스 라이선스[40]를 사용해야 함
- 따라서 국내 데이터의 저작권자들이 한국의 Creative Commons Attribution Licence 2.0 KR (CC BY 2.0 KR)[41]을 사용하기를 권장하는데 이 라이선스는 데이터의 저작자를 표시하기만 하면 데이터의 복제, 배포, 변형, 2차 저작물 작성과 상업적 이용도 가능하기 때문에 사용자들이 이해하기 쉽고 데이터도 보다 널리 이용될 수 있음
- 또한 상업적 이용을 금지하는 라이선스도 인정하지만 많은 데이터들이 서적 혹은 상업적으로 이용되는 저널에 콘텐츠로 포함되는 경우가 많으므로 상업적 이용을 금지하는 라이선스를 권장하지 않음
- 데이터 저작권자들이 전체 데이터에 대해서 단일 라이선스를 적용하기를 권장하나 때에 따라서는 각각의 개별 데이터에 대해 각각 다른 라이선스를 적용하는 것도 가능함
- 한편, 저작권이나 라이선스 문제는 학술 인용, 출처, 감사 표시 등과는 별개의 문제이며 국제적으로도 모든 데이터 이용자들은 특히 퍼블릭 도메인의 저작권이 없는 데이터를 학술적으로 이용하는데 있어서도 데이터 생산자 또는 기여자들을 바르게 인용할 것을 권유하고 있음

(2) 출판 비용

- 오픈 억세스 출판과 저장을 지원하기 위해 데이터를 출판하는 데이터 리파지토리 또는 데이터센터의 성격에 따라서 또는 데이터의 웹 서비스를 위해 데이터 출판자가 데이터를 가공하는 경우 데이터의 크기와 가공 난이도에 따라서 별도의 출판 비용이 소요될 수가 있음
- 따라서 데이터를 출판하고자 하는 데이터 저작권자는 사전에 데이터 출판자의

40) https://creativecommons.org/ (2017. 2.)
41) http://ccl.cckorea.org/about/ (2017. 2.)

데이터 정책을 잘 살펴보고 자신의 데이터가 웹에서 어떤 형태로 출판되고 영구 식별자의 부여 여부, 데이터의 관리 요령, 회원 제도 등을 체크하는 것이 좋음

3 데이터와 미디어 파일 제출

- 데이터를 제출하는 방법에는 여러 가지가 있으며 파일은 취급하는데 있어서 특별한 소프트웨어가 필요 없는 오픈 파일 포맷으로 변환하는 것이 좋음. 오픈 포맷은 저장하거나 다양한 컴퓨팅 플랫폼에서 쉽게 사용할 수 있음
- 파일 포맷은 컴퓨터 파일 안에서 정보를 조직화하는 방법을 기술하는 것이며 문서, 이미지, 오디오와 비디오 파일, 그리고 연구 데이터 셋에 적용시킬 수가 있음.
- 데이터를 생성하고 저장하는데 이용할 수 있는 파일 포맷에는 여러 가지가 있으며 데이터의 보존과 공유를 위해 적절한 파일 포맷을 선정함으로써 미래에도 데이터에 지속적으로 접근할 수 있고 데이터를 재활용할 수 있음
- 또한 표준 파일 포맷은 효율적인 데이터 공유에 가장 중요하며 많은 경우에 각 학문 분야별로 연구 데이터의 저장과 보존을 위해 선호하거나 필수적인 표준이 있는데 예를 들면 사회과학 데이터 셋의 경우 SPSS 데이터 파일을 들 수 있음

4 품질 심사, 편집과 표준 기입

- 데이터의 품질을 체크하기 위해서는 분류 등에 사용하는 용어의 일관성, 일반인이 이해할 수 있도록 약어나 수치 코드의 디코딩, 각 항목 입력 값의 오류 검증, 사용되는 식별자 등을 검토해야 함
- 데이터 편집자들은 데이터 테이블이나 미디어, 파일 등을 특정 데이터 구조를 통해 통합하며, 이러한 구조는 서로 다른 데이터 요소들을 통합, 검색, 열람, 분석할 수 있도록 함
- 이런 과정을 통해서 연구자들은 자신들의 데이터가 출판되기 전에 어떤 형태로 보이는지를 점검할 수 있으며, 이 과정에서는 추가 디버깅과 오류 수정, 용어와 기록 시스템의 문서화, 특정 데이터에 대한 저작권과 인용 관련 안내, 데이터의 활용성, 재이용 가능성에 대한 편집인의 의견 등이 포함됨
- 데이터는 상호운용성과 다른 컬렉션과의 통합, 새로운 연구 기회를 가능하게 하기 위해서 국제적 표준을 적용함

5 출판, 리파지토리 저장과 공개

(1) 출판

- 저자, 편집자들이 데이터가 출판을 위해 준비가 완료되었다고 판단하면 웹을 통해 일반에 공개하게 되는데 이 과정에서는 적당한 리파지토리를 통해 데이터를 저장하고 장기간 인용을 위해 필요한 영구 식별자를 부여하며 2차 배포 채널로서 GitHub[42]에 포스팅하거나 검색을 위한 색인과 함께 보다 널리 검색되도록 구글과 같은 검색엔진에 색인함

(2) 리파지토리 저장과 공개[43]

- 연구 데이터를 제출 받아 관리하며 장기적으로 보존할 수 있는 데이터 인프라를 갖추지 않았을 경우에는 데이터의 접근을 허용하는 데이터 센터나 데이터 리파지토리에 데이터를 제출하여 데이터의 공유와 보존을 확보하는 것이 좋음
- 이러한 리파지토리는 일반적으로 아래와 같은 요구 사항들을 가지고 있음
  . 주제 또는 연구 분야
  . 데이터의 재이용과 접근
  . 파일 포맷과 데이터 구조
  . 메타 데이터
- 또한 어떤 저널이나 학회들은 자신들의 데이터 셋을 장기 보존할 리파지토리에 대한 기본적 판단 기준을 공개하고 있음. 예를 들어 Earth System Science Data (Journal)의 리파지토리에 대한 판단 기준은 아래와 같음
  . 영구 식별자 : 데이터 셋은 DOI와 같은 디지털 콘텐츠 식별자를 반드시 가져야 함
  . 오픈 억세스 : 데이터 셋은 무료로 이용할 수 있어야 하고 로그인을 위한 무료 등록 절차를 제외하고는 데이터 접근에 어떤 장벽도 있어서는 안됨
  . 자유로운 저작권 : 원저자를 인용하는 한 누구나 데이터 셋을 무료로 복사, 배포, 전달, 활용할 수 있어야 하며 이는 Creative Commons Attribution License에 해당됨
  . 장기 이용 권한 : 리파지토리는 데이터 셋에 대해 장기간 이용 권한과 영구적 접근 권한을 보장해야 함
- 저자들은 그들의 데이터 셋을 제출할 데이터센터가 이러한 판단 기준을 충족하는지 사전에 면밀히 조사해야 하며 편집자들은 저자들의 데이터 리파지토리가 이 기준에 합당한지를 조사하고 만약 합당하지 않을 경우에는 데이터 (또는 데이터 저널) 출판 전에 저자와 이 문제를 상의해야 함

(3) 분야별 리파지토리

- 세계적으로 많은 리파지토리들이 re3data.org[44]나 OpenDOAR[45] 또는 다른 리파지토리 디렉토리에 등록되고 있으니 검색해 보길 바라며 여기에서는 분야별로 대표적인 데이터 리파지토리에 대해 소개함

---

42) https://github.com/ (2017. 2.)
43) https://library.uoregon.edu/datamanagement/repositories.html, (2016. 10.)
44) http://www.re3data.org/ (2017. 2.)
45) http://www.opendoar.org/ (2017. 2.)

- 화학
  . Cambridge Structural Database[46] : 작은 분자의 결정 구조
  . ChemSpider[47] : 화학 구조와 관련 정보를 수집한 것으로 무료 접근 가능
  . eCrystals[48] : X-선 결정 구조 데이터
  . PubChem[49] : 고분자가 아닌 저분자에 대한 생물 활성/바이오 에세이 데이터를 저장하고 있는 NCBI의 리파지토리로서 텍스트와 구조 기반의 검색 도구가 제공됨
- 컴퓨터 과학
  . Cooperative Association for Internet Data Analysis (CAIDA)[50] : 네트워크 기능의 과학적 분석을 위한 데이터 저장소
- 환경과 지질 과학
  . Marine Geoscience Data System (MGDS)[51] : 컬럼비아 대학 Lamont-Doherty Earth Observatory에서 주관하고 있는 NSF 지원의 해양 연구 프로그램을 위한 데이터 포털
  . National Climatic Data Center (NCDC)[52] : 기상과 고기후학
  . National Oceanographic Data Center (NODC)[53] : 전세계 해양 환경과 생태계 데이터
  . GEON[54] : 데이터 셋과 가시화 도구를 위한 포털
  . National Snow and Ice Data Center (NSIDC)[55] : 대지 필드 연구와 위성으로부터의 빙권 데이터 셋
- GIS와 지리정보
  . Geospatial Data[56] : 연방, 주, 지방 정부의 지리 데이터의 원스톱 서비스
  . GeoCommons.com[57] : GIS 파일 리파지토리와 검색 도구
  . Federal Geographic Data Committee[58] : 국가 공간 데이터 인프라 스트럭쳐 (NSDI) 클리어링 하우스 네트워크와 geodata.gov 포털에 접근성을 제공
  . National Geographic Data Center[59] : 데이터 셋의 저장소
- 생명과 생물과학
  . Dryad[60] : Dryad는 기초와 응용 생명과학의 심사 논문의 기초가 되는 데이터에 대한 국제적 리파지토리로서 노스 캐롤라이너 대학 메타데이터 연구센터

46) https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/ (2017. 2.)
47) http://www.chemspider.com/ (2017. 2.)
48) http://ecrystals.chem.soton.ac.uk/ (2017. 2.)
49) https://pubchem.ncbi.nlm.nih.gov/ (2017. 2.)
50) http://www.caida.org/home/ (2017. 2.)
51) http://www.marine-geo.org/index.php (2017. 2.)
52) https://www.ncdc.noaa.gov/ (2017. 2.)
53) https://www.nodc.noaa.gov/ (2017. 2.)
54) http://www.geongrid.org/index.php (2017. 2.)
55) https://nsidc.org/ (2017. 2.)
56) https://www.data.gov/geospatial/ (2017. 2.)
57) http://geocommons.com/ (2017. 2.)
58) https://www.fgdc.gov/ (2017. 2.)
59) https://www.ngdc.noaa.gov/ (2017. 2.)
60) http://datadryad.org/ (2017. 2.)

(University of North Carolina Metadata Research Center)와 국립 진화합성센터 (National Evolutionary Synthesis Center)에 의해 개발되었음

. Worldwide Protein DataBank[61] : 실험적으로 결정된 단백질 및 핵산과 같은 고분자 구조 정보이며 검색과 가시화 도구를 제공함

. UniProt[62] : 무료 단백질 시퀀스

- 물리학

. HEPData[63] : 수치 HEP 산란 단면적에 대한 고에너지 물리 반응 데이터베이스

. NIST Physical Standards Laboratory[64] : 물리 참조 데이터와 물성 테이블

. National Nuclear Data Center[65] : 핵구조, 핵반응, 핵반감기 데이터베이스가 포함되어 있음

- 사회과학

. ICPSR (Inter-university Consortium for Political and Social Research)[66] : 미시간 대학에 소재한 회원제의 비영리 데이터 저장소

. Dataverse Network[67] : 사회과학 분야의 연구 데이터를 dataverse라는 가상 데이터 저장소에 컬렉션 형태로 저장하고 있으며 IQSS (Institute for Quantitative Social Sciences at Harvard)에서 운영하고 있음. 이용자는 자신의 dataverse를 생성하고 데이터를 올릴 수 있음

(4) 리파지토리 디렉토리

- re3data ("REgistry of REsearch REpositories")[68] : 리파지토리의 목록
- DataCite[69] : British Library, BioMed Central, 그리고 UK's Digital Curation Centre에서 정리한 리파지토리의 목록
- Distributed Data Curation Center[70] : Purdue University 도서관에서 관리하는 데이터 리파지토리로서 과학 분야의 대한 50개 이상의 오픈 데이터 리파지토리 목록을 소개
- Gene Expression Omnibus (GEO)[71] : Gene Expression Omnibus (GEO)는 과학 학술단체에서 제출한 마이크로어레이, 차세대 시퀀싱, 그리고 다른 형태의 기능성 유전체 데이터에 접근할 수 있도록 하는 오픈 데이터 리파지토리
- Global Change Master Directory (GCMD)[72] : GCMD는 NASA에서 운영하는 지구과학 디렉토리로서 기후 변화와 지구과학 연구와 관련된 25,000건 이상의 지구와 환경 과학 데이터 셋에의 접근을 제공

61) http://www.wwpdb.org/ (2017. 2.)
62) http://www.uniprot.org/ (2017. 2.)
63) https://hepdata.net/ (2017. 2.)
64) https://www.nist.gov/pml/productsservices/physical-reference-data (2017. 2.)
65) https://www.nndc.bnl.gov/ (2017. 2.)
66) https://www.icpsr.umich.edu/icpsrweb/ (2017. 2.)
67) http://dataverse.org/ (2017. 2.)
68) http://www.re3data.org/ (2017. 2.)
69) https://www.datacite.org/ (2017. 2.)
70) https://www.lib.purdue.edu/researchdata (2017. 2.)
71) https://www.ncbi.nlm.nih.gov/geo/ (2017. 2.)
72) http://gcmd.nasa.gov/ (2017. 2.)

- MIT Data Management and Publishing[73] : MIT 도서관에서 발행하는 데이터 관리와 출판에 관한 지침서에는 천문, 대기과학, 생물학, 화학, 지구과학, 해양과 우주과학 분야의 오픈 데이터 리파지토리 목록을 제공하고 있음
- Oceanographic Data Repositories : West Coast Ocean Data Portal[74], National Oceanographic Data Center[75], Integrated Science Data Management[76], SeaDataNet[77], Integrated Ocean Drilling Program[78] 등이 있음
- Open Access Directory[79] : Simmons College의 도서관 정보과학 대학원에서 운영하며 2008년 시작된 데이터 리파지토리로서 Open Access Directory는 다학제 오픈 데이터 리파지토리뿐만 아니라 고고학, 생물학, 화학, 환경과학, 지질학, 지구과학, 지리공간 데이터, 해양과학, 의학, 물리학 분야의 50개 이상의 오픈 데이터 리파지토리에 링크를 걸어주는 wiki임
- Public Data Sets on Amazon Web Services[80] : 아마존 웹서비스는 공공영역과 비독점적 영역의 천문학, 생물학, 화학, 기후 분야에 대한 데이터 셋을 다운로드 할 수 있는 중앙 집중형의 데이터 저장소이며 한국어 웹사이트[81]가 있음

73) https://libraries.mit.edu/data-management/#0 (2017. 2.)
74) http://portal.westcoastoceans.org/ (2017. 2.)
75) https://www.nodc.noaa.gov/ (2017. 2.)
76) http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/index-eng.html (2017. 2.)
77) http://www.seadatanet.org/ (2017. 2.)
78) http://iodp.tamu.edu/curation/index.html (2017. 2.)
79) http://oad.simmons.edu/oadwiki/Main_Page (2017. 2.)
80) https://aws.amazon.com/public-datasets/?nc1=h_ls (2017. 2.)
81) https://aws.amazon.com/ko/public-datasets/ (2017. 2.)

IV DOI

1 DOI 시스템의 정의

- 데이터 셋을 식별하기 위해 사용되는 영구 식별자에는 Handles, Archival Resource Keys (ARKs), Persistent URLs (PURLs)와 같은 몇 가지 종류가 있으며 이들은 모두 인터넷 소재지를 해석하여 해당 콘텐츠가 소재하는 인터넷 소재지로 안내하는 것으로서 이 가운데 국제적으로 가장 범용적으로 사용되는 것이 DOI(Digital Object Identifier)[82]임
- 디지털 콘텐츠 식별자(DOI) 시스템은 주로 출판사들이 사용하는데 디지털 환경에 있는 지적 재산을 식별하기 위해 사용하는 것으로서 영구 식별자에 대한 핸들 시스템을 구현한 것임.
- 국제 DOI 연맹(IDF, International DOI Federation)[83]에서는 DOI 등록기관(RA, Registration Agency)들을 임명하는데, 이 DOI 등록기관들은 DOI의 접두사(prefix)를 할당하고, DOI 이름(DOI name)을 등록하며 등록자들이 메타데이터를 작성하고 유지하는데 필요한 시스템 인프라를 제공함
- DOI 시스템의 주요한 적용 분야는 아래와 같음
  . 약 3,000개 출판사들의 컨소시엄인 CrossRef를 통한 학술자료(저널 논문, 서적 등)의 영구 인용
  . 세계적인 연구 도서관, 과학기술정보기관, 과학 데이터센터들의 컨소시엄인 DataCite를 통한 데이터 셋의 인용
  . EU 출판국을 통한 유럽 연합의 공식 출판물의 인용

2 DOI명

- DOI명은 핸들 시스템이 식별할 수 있는 특정한 타입이며, 지적 재산 형태의 어떤 콘텐츠에 대해서도 부여할 수 있음. DOI는 '디지털 콘텐츠에 대한 식별자'가 아니라 '콘텐츠에 대한 디지털 식별자'라고 해석되어야 함
- DOI는 슬래쉬(/)로 구분되는 두 개의 부분, 즉 접두사(prefix)와 접미사(suffix)로 나누어지며, 독특하고 대소문자를 구별하지 않는 알파벳과 숫자로 이루어진 문자 배열로 구성됨
- 접두사는 DOI 등록기관에 의해 부여되며 항상 '10'으로 시작되는데 이것은 다른 타입의 핸들(Handle) 시스템과 DOI를 구별될 수 있도록 함. 접미사는 콘텐츠

82) http://ands.org.au/guides/doi.pdf, (2016. 9.)

83) http://www.doi.org/index.html (2017. 2.)

에 대한 정보를 제공하는 기관인 출판사들에 의해 부여되며 접두사 내에서 유일무이해야 함

- 예를 들어 ‘10.1594/PANGAEA.484677’의 경우, 접두사 10.1594는 디렉토리 코드 ‘10’(DOI 이름은 항상 10)과 등록자 코드 ‘1594’로 구성되어 있으며, 등록자 코드 1594는 DOI 등록기관인 German National Library of Science and Technology가 그 역할에 따라 데이터 셋에 대해 할당한 것임
- 상기 콘텐츠에 대한 메타데이터를 제공한 출판사는 Publishing Network for Geoscientific & Environmental Data이며 이 출판사는 해당 데이터 셋에 대한 유일무이한 식별자로서 접미사 ‘PANGAEA.484677’을 부여하였음. 상기 DOI에 대한 인용은 ‘doi:10.1594/PANGAEA.484677’와 같은 형태를 취해야 하나, 저자들은 데이터 셋으로 바로 링크가 될 수 있는 하이퍼텍스트 링크인 ‘http://dx.doi.org/10.1594/PANGAEA.484677’로 사용하는 것이 좋음
- 자신들의 데이터 셋에 DOI를 등록하기를 원하는 연구자 개인들은 일반적으로 연구 주제 분야별 데이터 아카이브(disciplinary data archive)나 기관 데이터 리파지토리(Institutional data repository) 또는 figshare나 Synapse와 같은 데이터 공유 서비스를 통해 자신들의 데이터 셋을 제출하고 DOI를 등록할 수 있음

3 DOI명과 다른 영구식별자

(1) DOI명과 다른 영구식별자의 차이점

- DOI명은 영구식별자(PID, Persistent Identifier)로서 많은 장점들을 가지고 있는데 예를 들면 디지털 콘텐츠, 비디지털 콘텐츠에 관계없이, 또한 콘텐츠들이 인터넷 상에 있든지 여부에 관계없이 이러한 콘텐츠들을 독특하게 식별할 수가 있음
- DOI명은 콘텐츠를 DOI 레지스트리에 기재함으로써 콘텐츠 자체를 영구적으로 식별하는데 반해 PID는 콘텐츠의 장소만을 영구적으로 식별함.
- DOI명은 국제 DOI 연맹(IDF)과 DOI 등록기관의 시스템 인프라에 의해 식별되며 이러한 시스템 인프라는 DOI 서비스들을 지속적으로 제공할 뿐만 아니라 DOI명의 품질과 정확성에 있어서 높은 신뢰도를 가지게 함
- 콘텐츠는 출판 과정을 거치면서 복수의 DOI들과 PID들을 가질 수가 있음. 만약 하나의 콘텐츠가 인터넷에 소재지를 가질 경우, 콘텐츠는 DOI 외에도 URL이나 Handle, PURL, ARK와 같은 다른 식별자를 가지게 되며, 각각의 DOI와 PID는 데이터 셋에 대해 각각 다른 장점들을 가지고 있음

(2) 데이터 셋에 DOI를 부여할 경우의 장점

- 국제 DOI 인프라를 통해서 DOI명을 부여하게 되면 관련 비용이 발생하게 됨. 따라서 DOI명은 데이터 셋의 내용이 적절하게 작성되고 장기적으로 접근할 수 있는 안정된 저장소에서 관리되는 데이터 셋에 한해서 공인된 기관들이 DOI명을 부여하는 것이 좋음

- 이는 DOI명을 가진다는 것은 해당 데이터 셋이 적절하게 관리되고 장기적으로 활용할 수 있도록 안정적으로 콘텐츠에 접근할 수 있다는 의미이며, 출판된 데이터는 출판계에서 1등급 연구성과물임을 의미하는데 그 이유는 해당 데이터 셋이 현존하는 학술 출판물에서 DOI를 부여받았기 때문임
- 이와 같은 방법으로 DOI를 사용하게 되면 인터넷에 존재하는 연구 데이터에 대한 접근성이 향상되고 과학 논문에서도 연구 데이터를 합법적으로 인용할 수 있는 콘텐츠로 인정하게 되며 연구 결과를 검증하고 미래의 연구에 다시 재활용할 수 있는 데이터 아카이빙 활동은 더욱 중요하게 됨

V 데이터 인용

1 데이터 인용의 정의

- 데이터 인용[84][85]이란 연구자들이 저널 논문이나 기술 보고서, 컨퍼런스 논문과 같은 연구 성과물에 참고 문헌으로 서지 정보를 관례적으로 인용하는 것처럼 데이터를 인용하는 것을 말함
- 데이터는 과거에 연구자들 사이에서 자주 공유되기는 했지만 저널 논문이나 다른 출판물과 같은 방법으로 인용되지는 않았으나, 최근 이러한 문화가 변하고 있는데 데이터를 인용한다는 것은 데이터를 연구의 부산물로 보기보다는 연구의 1차 성과물로서 인식하며 데이터 셋이 인용되면 학술 커뮤니케이션 사이클 안에서 데이터가 한층 검증되고 중요성을 가지게 됨
- 이는 데이터를 인용함으로써 기관이나 조직이 데이터 성과물에 대한 학술적 노력을 인정받고 보상받을 수 있는 가능성이 높아지게 되며, 또한 데이터는 출판물의 또 다른 형태인 학술적 성과물로서 활용될 수 있다는 것을 의미함
- 데이터에 디지털 콘텐츠 식별자(DOI)를 부여하면 데이터를 쉽게 인용할 수가 있어서 DOI는 데이터 인용에 가장 좋은 도구가 되며, DOI는 영구 식별자의 한 형태로서 데이터 셋에 부여하게 되면 데이터 셋이 적절하게 관리되고 장기적으로 사용될 수 있도록 접근성이 보장된다는 것을 의미함. 이는 현재 출판사들이 저널 논문에 DOI를 부여하고 저자들은 논문 인용에 이 DOI를 포함시키는 관례와 같다고 할 수 있음.
- 데이터의 인용에 있어서 최근의 동향은 아래와 같음
  . 데이터의 생성을 연구의 1차 성과물로서 인정하는 사례가 늘고 있음
  . 주제별 또는 기관별 데이터 리파지토리의 글로벌 네트워크가 발전하고 있는데 여기에서는 연구 데이터 컬렉션이 적절하게 작성된 메타데이터를 가지고 있으

84) http://ands.org.au/cite-data/index.html, (2016. 9.)
85) http://ands.org.au/guides/data-citation-awareness.pdf, (2016. 9.)

며 이는 영구 식별자를 통해 데이터 인용을 가능하게 함
. 정부의 연구비로 생산된 데이터에 대한 접근성이 향상된다는 것은 보다 많은 연구 데이터가 공개되고 재사용 된다는 것을 의미하며 따라서 데이터가 바르게 인용되고 계량, 추적될 수 있음
. 데이터 인용 정보는 향후에 연구자의 연구 평가와 보상에 반영될 것임
. EndNote와 같은 서지정보 관리시스템이 현재 연구데이터 인용을 위한 템플릿을 포함하고 있음

2 데이터 인용과 출판물에의 연계 이유

- 데이터를 인용하려는 이유[86]는 연구 과정에서 생산된 데이터가 현재의 학술 논문 또는 단행본이 가지는 학술적 가치와 동일한 가치를 가진다는 인식 때문이었음
- 전통적으로 학술 잡지는 지식을 전파함으로써 연구활동을 지원했는데, 예를 들면 과학자들은 연구 목적과 방법에 기초하여 도출한 결론을 심사하고 이 결론을 근거하여 다시 새로운 연구를 수행하는 방식으로 지식을 생산하고 이를 학술 잡지를 통해 전파하였음. 그러나 많은 학문 분야에서 이제는 논문만으로는 이런 목적을 달성할 수가 없으므로 그 저변에 있는 데이터까지 공유할 필요성을 느끼게 되었음
- 매체로서 저널 논문이 성공을 거둔 이유는 검증 시스템 덕분이라 할 수 있는데, 이 시스템은 저자들이 자신의 연구를 공개하고 공정한 심사를 통해 명성을 얻게 하는 체계와 이러한 연구 노력을 계량적으로 측정하여 저자와 소속기관에게 보상하며 이러한 연구 결과를 지속적으로 인용하고 재사용할 수 있는 저장소가 있었기 때문임
- 만약 데이터가 연구의 1급 기록물로서 간주되기 위해서는 데이터에도 상기의 논문 검증 시스템과 유사한 검증 시스템이 구축될 필요가 있으며 이를 위해 전통적인 출판물 내에서 데이터를 인용하는 강력한 인용 체제가 정착되어야 함
- 인용에는 신뢰할 수 있는 데이터에 대한 책임자의 이름이 포함되어야 하고 국제적으로 통용되는 고유 식별자를 데이터에 부여함으로써 개별 데이터의 영향을 측정할 수 있어야 함. 또한 인용을 통해서 데이터의 위치와 접근에 필요한 정보들을 제공할 수가 있으며, 이러한 출판물의 인용 방식을 사용함으로써 데이터는 저널 논문을 관리하는 기존의 인프라의 이점을 그대로 활용할 수가 있게 됨
- 전자저널 서비스는 개별 논문들의 순위를 매길 수 있고, 현재의 논문을 인용하고 있는 다른 논문으로 연결되도록 함으로써 독자들이 개별 논문의 영향을 추정할 수 있도록 함.
- 또한 해당 논문을 자신의 논문에 인용하며, 어떤 경우에는 다른 연구자들이 발견

86) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)

한 오류나 문제점들을 파악할 수 있는 이점이 있음. 같은 방법으로 데이터에서 해당 데이터를 인용하고 있는 논문으로 연결하게 하면 전자 저널 서비스가 갖는 동일한 이점을 누릴 수 있을 뿐만 아니라 데이터의 내용을 설명하는 문서도 함께 읽을 수가 있음

- 만약 과학과 연구 커뮤니티 전체의 문화가 데이터 공유라는 방향으로 이동하고 과학의 신속성과 투명성이 증대되어 과학이 진보하게 되면 결국 데이터와 논문의 서지적 연결은 필요한 절차로 정착할 것으로 생각됨

3 데이터 인용의 이점

- 데이터를 인용함으로써 연구자와 연구기관들에게 가져올 이점으로 생각되는 것은 아래와 같음
  . 관련 출판물에 인용할 수 있는 데이터를 포함하게 되면 해당 출판물의 인용 빈도를 향상시킨다는 보고가 있음[87].
  . 데이터의 인용을 관례화하게 되면 연구자들이 데이터를 1급 연구 성과물로서 인정할 수 있게 됨
  . 본인의 이력서나 연구실적에 본인의 저널 논문과 기술보고서, 컨퍼런스 논문과 함께 데이터의 출판과 인용 내용도 기재할 수 있음
  . 저널 논문의 인용지수 측정과 유사한 방법으로 인용된 데이터에 대해 그 영향을 측정하기 위한 추적과 계량화가 가능함
- 데이터는 DOI가 없이도 인용될 수 있지만 데이터에 DOI를 부여하고 데이터 인용에 이 DOI를 포함하게 되면 좋은 활용사례가 될 수 있으며 DOI는 아래와 같은 부가적인 이점을 가져올 수 있음
  . 인터넷을 통해 연구 데이터에 쉽게 영구적으로 접근하여 이용할 수 있음
  . 데이터를 재이용하고 연구결과를 검증할 수 있도록 데이터의 열람, 검색, 관리 기능이 향상됨
  . Thomson Reuters Data Citation Index와 같은 인덱싱 서비스에 의해 데이터의 자동 추적이 가능하며 열람 횟수, 다운로드 횟수 등을 파악할 수 있음

4 데이터 인용 원리

- 데이터 인용 원리[88]는 인용의 목적, 역할, 기여에 대한 것으로서 인간이 이해할 수 있고 시스템이 실행할 수 있는 인용 사례를 만들 필요성이 있어서 작성되었음.

87) http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000308, (2016. 9.)
88) https://www.force11.org/datacitation, (2016. 9.)

- 데이터 인용 사례는 학제 간에 매우 다양하고 기술은 시간에 따라 발전하기 때문에 이러한 인용 원리는 특정 분야를 위한 권고사항이 포함되어 있지 않으며 오히려 이러한 인용 원리들을 구현하기 위한 사례와 도구들을 개발하도록 커뮤니티를 지원하기 위한 것임
- 이러한 원리들은 중요성으로 판단하기보다 쉽게 이해할 수 있도록 아래와 같이 분류하였음

① 중요성 : 데이터는 합법적이고 인용 가능한 연구성과물로서 인식되어야 하고 다른 학술 출판물의 인용처럼 데이터 인용도 연구논문 안에서 동일한 중요성을 가져야 함
② 신뢰와 기여 : 데이터 생성에 다양한 방법으로 기여한 기여자는 데이터의 인용을 통해 학술적으로 신뢰와 인정을 받아야 함
③ 증거 : 학술 문헌에서 주장의 논리는 때와 장소에 상관없이 데이터에 근거해야 하며 해당 데이터는 반드시 인용되어야 함
④ 고유 식별 : 데이터 인용은 시스템이 실행할 수 있고 국제적이며 학술 커뮤니티에서 널리 사용되는 영구적으로 식별할 수 있는 방법을 포함해야 함
⑤ 접근 : 데이터 인용은 이러한 참고 데이터를 인간과 시스템이 모두 사용할 수 있도록 데이터 자체와 관련 메타데이터와 문서, 코드와 기타 자료들에 쉽게 접근할 수 있도록 해야 함
⑥ 영속성 : 데이터와 해당 데이터의 저장소를 설명하는 고유 식별자와 메타데이터는 해당 데이터의 생애 기간이 지난 뒤에도 영속되어야 함
⑦ 특정과 검증 : 데이터의 인용은 주장 논리를 설명하는 특정한 데이터에 대해 쉽게 식별, 접근, 검증할 수 있어야 함. 인용 또는 인용 메타데이터는 검색된 데이터의 특정 시간과 버전, 데이터의 단위 등이 원래 인용된 것과 동일하다는 것을 검증하기에 충분한 출처 정보를 포함하고 있어야 함
⑧ 상호 운용성과 유연성 : 데이터의 인용방법은 학제 간의 다양한 사례를 수용할 수 있도록 유연해야 하며 서로 간에 큰 차이가 없어서 학제 간의 데이터 인용 사례에서 상호운용성을 확보할 수 있어야 함

5 데이터 인용 방법

- 출판물과 직접 관련이 있는 데이터를 인용하는 방법은 일반적으로 데이터에 대한 접근 방법을 알려주는 것임. 공개된 데이터에 대해서는 어떤 리파지토리에서 무엇을 이용할 수 있는지, URL, 식별자, 또는 데이터에 접근할 수 있는 접근 코드 등을 알려주어야 하며, 접근이 제한된 데이터에 대해서는 접근을 제한하는 법적 또는 도덕적 이유를 설명하고 접근 조건을 설명하는 문서에 영구적으로 연결이 되도록 해야 함
- 그러나 이런 종류의 단순한 설명으로 데이터를 인용할 수는 있으나 아래와 같은

몇 가지 부족한 측면이 나타날 수가 있음.

. 만약 식별자나 URL에 인쇄상의 오류가 있을 경우에는 리파지토리에 보관된 데이터 자원 중에서 해당 데이터를 찾을 수 있는 추가적인 정보가 없음

. 저자들은 특정 데이터를 지정하기보다는 리파지토리의 URL을 알려주는 경향이 많음

. 데이터 생성자에게 주는 보상이 없으며 특히 데이터 생성자와 출판물의 저자가 다를 경우에는 더욱 중요한 문제가 됨

. 데이터를 연구의 1등급 기록물로 취급하지 않음

- 위의 모든 문제들은 데이터 인용으로 해결될 수 있으며 이러한 데이터 인용은 다른 문헌 인용과 동일하게 참고 문헌에서 인용되어야 함
- 그러나 출판사들이 데이터 인용을 꺼려하는 경우에는 참고 문헌에 데이터 논문을 인용하는 방법을 생각해 볼 수 있음. 이 데이터 논문은 데이터와 수집 방법만을 기술하며 데이터로부터 과학적 결론을 이끌어 내지는 않는데, 이런 데이터 논문은 일반 학술저널의 특집으로 출판되거나 Earth System Science Data와 같은 데이터 저널에 데이터 논문으로서 출판됨
- 논문의 본문 내에서 데이터에 대한 접근 방법을 기술하는 형태의 데이터 인용은 저널마다 다른데, 예를 들어 PLoS와 Pensoft에서는 “Data resources” 또는 “Data access and terms of use” 와 같은 특별 섹션을 사용하고 있으며 또 다른 저널에서는 초록 끝에 넣으라고 권유하기도 함
- 특별한 언급이 없는 경우에는 일반적으로 감사의 글에 데이터와 관련된 내용을 함께 언급하는데 그 이유는 연구비에 대한 감사의 글을 첨부하는 것이 연구비 지원에 대한 필수 조건인 경우가 많으며 이 필수 조건을 준수하는지 여부를 검토할 때에 데이터 관련 내용도 함께 읽을 수 있기 때문임. 또한 어떤 기관들은 데이터 접근에 대한 상세한 지침서를 제공하기도 함
- 또한 데이터로부터 다른 데이터를 인용하는 것도 가능한데 이렇게 함으로써 먼저 생성된 데이터를 사용할 수 있도록 유도하여 데이터의 공유를 향상시킬 수가 있음. 이를 위한 좋은 방법으로서는 데이터 셋에 테이블을 포함시킬 수가 있는데 이 테이블에는 원천 데이터 셋과 서브 데이터 셋의 리스트가 설명되어 있음. 다른 방법으로는 데이터를 설명하는 메타데이터에 다른 서브 데이터 셋과의 관계를 설명하는 항목을 포함시키는 것임

6 데이터 인용 요소

- 데이터 인용도 아래와 같이 다른 인용과 동일한 요소를 포함하고 있으며, 주제별 또는 기관별 리파지토리나 데이터센터에서는 일반적으로 데이터 인용을 위한 지침을 발표하고 있음
- 다만 이러한 지침이 없을 경우에는 서적을 인용할 때와 같은 형식을 따르는 것이 좋으며 데이터 인용을 위해 고려해야 하는 인용 요소들을 보다 자세히 소개

하면 아래와 같음[89]

. 저자 (author) : 데이터 셋을 생성한 생성자로서 개인이 될 수도 있고 개인이 모인 그룹 또는 기관이 될 수도 있음
. 출판 일시 (publication date) : 데이터가 출판된 연도나 데이터 셋이 온라인에 올라온 시기를 의미하며 가급적 최근 날짜로 함. 이 출판 일시는 데이터의 품질 검증 절차가 끝나고 데이터 셋을 이용할 수 있게 된 날짜나 엠바고 기간이 설정된 경우 엠바고 기간이 종료되어 이용이 가능하게 된 날짜를 뜻함
. 제목 (title) : 인용된 데이터 셋 자체의 이름으로서 데이터 시설 이름이나 데이터 셋이 들어 있는 콜렉션 이름을 포함할 수도 있음
. 편집 (edition) : 데이터 셋의 가공 정도를 나타내는 데이터의 처리 단계 또는 수준을 의미함
. 버전 (version) : 데이터 셋의 버전 정보를 의미하며 데이터를 추가하거나 데이터의 유도 과정을 재실행 하면 버전의 숫자가 증가함
. 자원 유형 (resource type) : 예를 들면 'database'나 'dataset'을 뜻함
. 출판자 (publisher) : 데이터를 보유하거나 품질 관리를 수행하는 기관을 뜻함
. 식별자 (identifier) : 지속적인 스키마와 일치하는 데이터에 대한 식별자를 뜻함
. 소재지 (location) : 데이터 셋을 이용할 수 있는 영구적인 URL로서 어떤 식별자 스키마는 식별자 리졸버 서비스를 통해 이 정보를 제공함

- 상기의 인용 요소 가운데에서 모든 데이터 인용에 공통으로 사용되는 중요한 요소로서는 저자, 제목, 출판 일시, 출판자와 소재지로서 이러한 정보들은 이용자들로 하여금 데이터 셋의 관련성을 판단하고 데이터 셋에 접근할 수 있도록 하며 데이터의 품질 혹은 지속성에 대한 확신을 줄 수가 있음
- 이것은 형식적으로 데이터 셋을 확인할 수 있으나 실제로는 범용적인 식별자가 필요한데 가장 효율적인 해결책은 DOI와 같이 식별자와 리졸버 서비스로 이루어진 데이터에 접근할 수 있는 소재지 정보를 제공하는 것임
- 데이터 인용에서 이러한 인용 요소들이 어떻게 결합되어 사용되는지는 텍스트 출판물의 인용을 근거로 하는데 아래에는 일반 형식의 데이터 인용의 사례와 데이터 리파지토리가 제시하는 데이터 인용 형식을 나타내었음

<일반 형식의 데이터 인용의 예>

APA

Cool, H. E. M., & Bell, M. (2011). Excavations at St Peter's Church, Barton-upon-Humber [Data set]. doi:10.5284/1000389

Chicago

(Footnote) H. E. M. Cool and Mark Bell, Excavations at St Peter's Church, Barton-upon-Humber (accessed May 1, 2011), doi:10.5284/1000389.

89) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)

(Bibliography) Cool, H. E. M., and Mark Bell. Excavations at St Peter’s Church, Barton-upon-Humber (accessed May 1, 2011). doi:10.5284/1000389.

MLA
Cool, H. E. M., and Mark Bell. “Excavations at St Peter’s Church, Barton-upon-Humber. “Archaeology Data Service, 2001. Web. 1 May 2011. <http://dx.doi.org/10.5284/1000389>.

Oxford
Cool, H. E. M. and Bell, M. (2011), Excavations at St Peter’s Church, Barton-upon-Humber [dataset] (York: Archaeology Data Service), doi:10.5284/1000389.

<데이터 리파지토리가 제시하는 데이터 인용 형식의 예>
PANGAEA
Willmes, S et al. (2009): Onset dates of annual snowmelt on Antarctic sea ice in 2007/2008. doi:10.1594/PANGAEA.701380

Dryad
Kingsolver JG, Hoekstra HE, Hoekstra JM, Berrigan D, Vignieri SN, Hill CE, Hoang A, Gibert P, Beerli P (2001) Data from: The strength of phenotypic selection in natural populations. Dryad Digital Repository. doi:10.5061/dryad.166

Dataverse
Frederico Girosi; Gary King, 2006, ‘Cause of Death Data’, http://hdl.handle.net/1902.1/UOVMCPSWOL UNF:3:9JU+SmVyHgwRhAKclQ85Cg==IQSS Dataverse Network [Distributor] V3 [Version].

- 데이터 인용 표준은 분야별 또는 출판사별로 다양하나, DataCite는 다음과 같은 형식을 추천하고 있음

생성자 (출판년도) 제목 출판자 식별자
Hanigan, Ivan. (2010) : Meteorological Data for Australian Postal Areas. Australian Data Archive. DOI : 10.4225/13/50BBFCFE08A12

- 또한 2개의 선택 항목(버전과 자원 형태)이 포함될 수도 있음

생성자 (출판년도) 제목 버전 출판자 자원 형태 식별자

. Version (Edition)이 추가될 경우

Colley, Sarah. (2010) Archaeological Fish Bone Images Archive Tables. 1st edition. Sydney eScholarship Repository Sydney.
http://ses.library.usyd.edu.au/handle/2123/6253

. Resource Type이 추가될 경우

Abraham, G; Kowalczyk, A; Loi, S; Haviv, I; Zobel, J. (2011) Computational Model for Gene Set Analysis to predict breast cancer prognosis based on microarray gene expression data.
Computer Science and Software Engineering, The University of Melbourne.
Computational Model. doi : 10.4225/02/4E9F69C011BC8

7 기여자 식별자

- 데이터의 생산자 또는 기여자가 일반적인 이름을 가지고 있거나 소속 기관이 자주 변경될 경우 이들을 정확하게 식별하기가 쉽지 않으며, 이를 해결하기 위해서는 개개인의 이름에 고유한 식별자를 부여하고 이를 개개인이 출판한 논문, 데이터 등과 연결시키는 것임[90]
- 이와 관련하여 이미 몇 가지 식별자 스키마가 잘 구축되어 있지만 대부분이 범위가 너무 좁거나 저작권이 있거나 개인의 식별이 아닌 인증을 위해 사용하고 있어서 개인 식별에는 만족스럽지 못함. 그러나 현재 특별히 개인 식별을 위해 개발되고 있는 것으로서 아래와 같은 두 가지 스키마가 있음
- Open Researcher and Contributor Identifier (ORCID)[91]는 특히 학술 저자들을 위해 만들어진 스키마이며, 현재 국제적인 대형 학술 출판사를 포함하여 300개 이상의 기관들로부터 지원받고 있으며 수많은 검색 시스템에 적용되고 있음. 연구자들은 자신들의 ORCID 프로파일에 학력과 경력 사항, 지원 받은 연구비뿐만 아니라 수행하고 있는 과제 리스트를 기재할 수 있으며, ORCID 프로파일은 Thomson Reuters' ResearcherID, Scopus, Scholar Universe, RePEc와 같은 다른 스키마의 식별자와 프로파일에 연결될 수 있음
- International Standard Name Identifier (ISNI)[92] 스키마는 ISO 표준으로서 지적재산권의 생성 또는 유통과 관련된 사람, 아호, 인격체, 법인과 같은 일반적인 이름의 식별자를 등록하는 것임. 따라서 ORCID보다 넓은 범위의 스키마이며 개인뿐만 아니라 기관도 식별할 수 있음. ISNI는 16자리의 숫자(마지막 자리는 X이지만)의 형태를 취하고 있으며 각각의 식별자는 이름, 생년월일, 활동 분야나 역할, 생성물의 제목, 추가 정보를 얻기 위한 URI 등 상세한 내용을 포함하는 메타데이터

90) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)
91) https://orcid.org/ (2017. 2.)
92) http://www.isni.org/ (2017. 2.)

를 가지고 있음
- 이러한 식별자의 1차적인 활용은 소프트웨어 도구를 지원하는 일이므로 사람이 검증하는 서류에 기재하기보다는 기계가독형 메타데이터에 기재하는 것이 좋음. 따라서 저자들은 ORCID 식별자나 이와 유사한 것들을 자신들의 참고 문헌 리스트에 포함시키려 하지 말고 논문이나 데이터를 출판사 또는 리파지토리에 제출할 시점에서 저자 자신들의 ORCID 식별자를 기재하는 것이 좋음

8 입도

- 인쇄 출판물에서는 다양한 입도를 가진 것들을 인용하는 것이 그리 큰 문제가 되지 않는데, 단일 저자의 단행본은 책 전체가 참고 문헌이 되며 저널의 경우에는 관련 논문이 개별적으로 참고 문헌이 됨
- 그러나 데이터 셋은 조금 더 복잡한 양상을 보이는데 일반적으로 데이터 셋은 몇 개의 테이블 또는 많은 데이터 포인트를 포함하는 파일들로 이루어져 있으며 이들은 콜렉션의 한 부분을 구성함. 또한 기능이나 파라메터와 같이 추상적인 데이터가 서브 셋 형태로 사용되는 수도 있음[93)]
- 저자에게 있어서 가장 실용적인 해결책은 입도 수준과 관계없이 식별자를 부여하기 위해 리파지토리가 선정한 모든 데이터 셋을 목록화 하는 것임. 만약 더 작은 입도의 데이터를 설명할 필요가 있는 경우에는 저자는 원문 속의 인용에서 서브 셋 데이터를 발견하는데 필요한 추가 정보를 독자들에게 제공해야 함. 이를 위한 규칙은 아직 만들어지지 않았으나 만약 리파지토리가 서로 다른 입도의 데이터에 대해 식별자를 부여할 경우에는 추가 정보를 줄이기 위해 인용의 요구조건을 충족시킬 수 있는 가장 작은 입도 수준의 데이터를 참고문헌 목록에 사용해야 함

9 공개되지 않은 데이터의 인용

- 아직 공개되지 않은 데이터 셋을 인용할 경우에 일반적인 규칙은 이미 알려진 정보는 최대한으로 참고 문서에 제공하는 것이며, 여기에는 최소한 데이터 셋의 생성자와 제목은 포함되어야 함[94)]
- 아직 데이터 셋이 저장되지 않았다면 수집 날짜를, 데이터 셋은 저장되었으나 아직 온라인에서 이용이 불가능할 경우에는 “출판 중” 이라는 날짜와 함께 출판사인 리파지토리 이름을 기입해야 함. 온라인 이용이 가능하게 되었을 경우에는 데이터 셋의 상태에 대한 상세한 내용(저장 여부, 엠바고 기간, 접근 제한, 완전 공개)이 데이터 접근 안내 정보에 상세히 기술되어야 함

93) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)
94) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)

- 아직 출판되지 않은 원고의 참고 문헌에서 공개되지 않은 데이터를 인용할 경우에 저자는 원고를 출판하기 전에 미공개 데이터 관련 정보가 최신의 것으로 갱신되었는지 여부를 참고 문헌에서 재차 확인해야 함

10 물리적 데이터(physical data)의 인용

- 샘플이나 물질과 같은 물리적 데이터의 인용 방법[95]과 디지털 데이터의 인용 방법 사이에는 원칙적으로 큰 차이가 없음. 물리적 데이터는 디지털 데이터에 비해 재생산과 공유가 어려운 측면이 있으나 디지털 데이터 역시 인터넷으로 공개하기에는 너무 민감하거나 용량이 큰 데이터의 전송에 대한 문제들이 있음
- 실제로 가장 혼란을 일으키기 쉬운 문제는 물리적 데이터에 대한 URL의 제공 여부와 제공 방법에 대한 것인데, 만약 물리적 데이터가 리졸버 서비스와 함께 식별자를 가지고 있다면 이 식별자는 URL에 사용되어야 함. 예를 들면 International Geo Sample Number(IGSN)는 카달로그에 기재되고 리졸버 서비스 URL인 "http://www.geosamples.org/profile?igsn=" 에 이 번호를 추가함으로써 물리적 데이터와 관련된 정보에 접근할 수 있게 됨
- 만약 물리적 데이터가 리졸빙 되지 않는 식별자를 가지고 있을 경우에는 참고문헌에서 이 사실을 언급해야 하며 URL은 데이터에 접근할 수 있는 방법을 설명하고 있는 페이지로 안내해야 함

11 동적 데이터(dynamic data)의 인용

- 전통적인 출판물과는 달리 어떤 데이터 셋들은 매우 동적인데 예를 들면 시간이 경과함에 따라 꾸준히 새로운 데이터가 추기된다든가 혹은 데이터 셋이 계속 갱신되고 수정되는 동적인 데이터들이 있음.
- 이에 대해 많은 연구 그룹들이 동적인 데이터 셋의 특별 버전을 인용하는 방법에 대해 아래와 같이 제안하고 있음[96]

① DataCite : Starr & Gastl (2011)[97]

. DataCite는 데이터 셋의 버전이 바뀔 때마다 데이터 셋을 매번 재등록해야 한다는 검증 규칙을 강요하진 않으나, 이 검증 규칙은 이런 특성의 데이터 셋의 인용 방법으로는 권장할 수 있는 최선의 사례라고 생각하고 있음

. DataCite 메타데이터 스키마는 버전 넘버를 포함하며 버전, 변수 등 다른 콘텐츠와의 다양한 관계를 정의할 수 있음

② Dataverse : Altman & Crosas (2013)

95) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)
96) http://datapub.cdlib.org/datacitation/, (2016. 9.)
97) http://www.dlib.org/dlib/january11/starr/01starr.html, (2016. 9.)

. 데이터 셋의 모든 버전은 하나의 DOI를 가지나 각 버전별 데이터는 별개의 버전 넘버를 가지고 인용됨
. 이전 버전의 데이터는 Dataverse 랜딩 페이지를 통해 접근할 수 있음

③ Digital Curation Center (DCC) : Ball & Duke (2012)[98]
. 신규 버전은 신규 식별자를 가지는 신규 데이터 셋임
. 데이터가 갱신되지 않고 시간 경과에 따라 단순히 추가되는 경우에는 추가된 데이터 셋에 대해서는 신규 식별자가 부여되며, 이용자는 버전별 데이터를 통합하여 전체 데이터 셋을 얻을 수 있음
. 갱신된 데이터 셋은 새로운 식별자가 부여되어 인용에 사용할 수 있으며 갱신 전의 데이터 셋도 그대로 보존됨

④ Earth Science Information Partners (ESIP)[99]
. 시간 경과에 따라 단순히 추가된 데이터 셋은 새로운 버전이 아니며 신규로 식별자를 받을 수 없음. 인용에는 접근 가능 날짜나 분석된 시간 범위와 같은 내용이 포함될 수 있음
. 데이터 셋이 갱신된 경우, 데이터 편집자는 버전의 중요도 여부를 반드시 구별해야 하며 각각의 버전의 특성과 파일/기록 범위를 기술해야 함. 예를 들면 전체 데이터의 재가공과 같이 전체 데이터 셋에 영향을 미치는 것은 중요한 버전으로 간주해야 함
. 중요한 버전은 신규 식별자와 콜렉션 레벨의 메타데이터를 가지게 되며 이전 메타데이터는 신규 버전을 알려주고 이전 버전의 데이터 상태를 설명해야 함
. 중요하지 않은 새로운 버전은 파일 레벨의 메타데이터 문서에 설명함

⑤ Lawrence. et al. (2011)[100]
. 데이터가 갱신되든지 단순히 추가되든지 상관없이 변경이 있을 경우는 모두 신규로 메타데이터가 작성되고 신규 식별자를 부여받음

⑥ Natural Environment Research Council (NERC) : Callaghan (2012)
. 모든 변경은 신규 버전과 신규 식별자를 부여받음

⑦ Organization for Economic Cooperation and Development (OECD) : Green (2009)
. 데이터의 갱신, 추가, 유지에 상관없이 랜딩 페이지로 인도하는 DOI를 인용하며 가능하다면 위키피디아와 같은 버전을 사용함

⑧ ZooKeys : Penev et al. (2009)[101]
. 정적인 데이터 테이블(예를 들면 스프레드 시트)과 역동적인 데이터베이스들을 구별함

98) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)
99) http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Detailed_Citation_Content, (2016. 9.)
100) http://projects.iq.harvard.edu/datacitation_workshop/files/lawea_datapublication_submitted.pdf#page=1&zoom=auto,0,815, (2016. 9.)
101) http://zookeys.pensoft.net/articles.php?id=1992, (2016. 9.)

. 데이터 테이블은 DOI를 부여하며 바뀌지 않음
. 데이터베이스의 인용은 버전, 날짜, 접근 가능 시각 등 주의 깊게 검토해야 할 사항들을 포함하여야 함

12 심층 인용(deep citation)

- 만약 출판 논문이 데이터 셋의 일부분만 사용하고 있다면 해당 데이터 셋의 서브 셋 데이터까지 상세한 데이터를 인용에서 사용해야 함.
- 데이터 셋이 구조적으로 크게 변화할 수 있고 데이터 생성자들은 미래의 이용자들의 요구를 충분히 예측할 수 없기 때문에 일반적인 해결책을 찾기는 어려우나 아래와 같이 몇 가지 사례에서 문제 해결을 위한 노력들을 찾아볼 수 있음

① DataCite Starr & Gastl (2011)[102]
. DataCite 메타데이터 스키마에는 IsPartOf, HasPart와 같이 다른 콘텐츠와의 다양한 관계를 정의할 수 있는 항목이 있음
. 데이터 셋의 서브 셋에 신규 DOI를 부여하고 메타데이터를 통해서 그 둘을 연계함으로써 데이터의 입도 문제를 해결할 수 있음

② Dataverse : Altman & King (2007)[103]
. 모든 데이터 셋을 인용하고 텍스트를 통해 서브 셋 데이터가 어떻게 생성되었는지를 설명함
. 서브 셋 데이터 생성이 간단한 경우에는 인용에서 이를 설명함
. 인용에서 서브 셋의 UNF(Universal Numerical Fingerprint)를 항상 포함시키도록 하는데 이는 두 서브 셋이 동일하다는 것을 확인하기 위해 사용되는 것임

③ Digital Curation Center (DCC) : Ball & Duke (2012)[104]
. 식별자를 인용에 사용하고 서브 셋 데이터가 어떻게 생성되었는지를 문서로 기술함

④ Earth Science Information Partners (ESIP)[105]
. 일반적인 정책은 없으며 데이터 운영자는 데이터의 서브 셋 데이터를 인용하는 방법을 제시해야 함
. 만약 서브 셋 데이터의 생성이 단순하다면 인용에 포함시킴

⑤ Lawrence. et al. (2011)[106]
. 가능하다면 인용에서 서브 셋 데이터를 설명하기 위해 정의된 레지스트리로

102) http://www.dlib.org/dlib/january11/starr/01starr.html, (2016. 9.)
103) http://www.dlib.org/dlib/march07/altman/03altman.html, (2016. 9.)
104) http://www.dcc.ac.uk/resources/how-guides/cite-datasets, (2016. 9.)
105) http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines#Detailed_Citation_Content, (2016. 9.)
106) http://projects.iq.harvard.edu/datacitation_workshop/files/lawea_datapublication_submitted.pdf#page=1&zoom=auto,0,815, (2016. 9.)

부터 정의된 콘텐츠 이름(예를 들면 콘텐츠 이름, 레지스트리의 URI)을 사용함

⑥ National Snow & Ice Data Center[107]

. 인용에서 서브 셋 데이터(예를 들면 시간적 또는 공간적 범위)를 기술함

13 데이터 인용에 대한 저자들의 의무

- 만약 학술 출판물의 증거로 사용하기 위해 데이터를 생성, 또는 수집했다면 가능한 빠른 시일 내에 해당 데이터들을 적당한 데이터 아카이브나 리파지토리에 제출하여 저장되도록 해야 함. 또한 데이터 아카이브나 리파지토리가 저장된 데이터에 대해 영구 식별자나 URL을 제공하지 않을 경우 이를 제공하도록 요청해야 함
- 논문에서 데이터 셋을 인용할 때에는 편집자나 출판사가 요구하는 인용 형식을 준수하며, 만약 특별한 인용 형식이 없을 경우에는 표준적인 데이터 인용 형식을 사용하고 서적 출판물의 인용 형식과 어울리도록 함
- 특별히 다른 것이 있지 않는 한, 데이터 셋의 식별자는 어디든지 가능한 URL 형식으로 부여함
- 서적 출판물의 참고 문헌에 데이터 인용을 포함시키며 시판 중인 참고문헌 관리 패키지 가운데에는 데이터 셋을 지원하는 것이 있으므로 이를 활용하는 것이 좋음
- 저자의 요구를 충족시킬 수 있는 가장 작은 입도의 데이터 셋을 인용하도록 하며 이 입도가 충분치 않을 경우에는 원문 속에서 이러한 서브셋 데이터에 대해 상세히 소개함
- 데이터 셋에 몇 가지 버전이 존재한다면 사용하는 정확한 버전을 인용하도록 함
- 데이터 셋을 인용하는 논문을 출판할 때에는 데이터 셋을 보유하고 있는 리파지토리를 알려줌으로써 데이터 셋과 이를 인용하는 논문이 서로 연결되도록 함

14 데이터 인용 지수

- Thomson Reuters Data Citation Index와 같은 인용 지수를 활용하면 연구 데이터의 재사용 횟수를 측정할 수가 있음[108]. 이것은 저널 논문이나 기타 학술 출판물의 인용을 측정하는 Web of Science and Scopus와 같은 제품을 사용하는 것과 유사한데 이러한 인용 지수의 측정은 성과 분석과 보고에 주로 사용됨
- 데이터 인용 지수는 연구 과정 전반에 걸쳐 관련된 모든 사람들이 데이터를 발견, 재이용하고 해석하는데 도움을 줄 수 있는데 구체적으로는 다음과 같음

107) http://nsidc.org/about/use_copyright.html, (2016. 9.)
108) http://wokinfo.com//products_tools/multidisciplinary/dci/about/, (2016. 9.)

<연구자들의 이점>
. 하나의 데이터 인용을 통해 가장 영향력 있는 리포지토리, 데이터 세트 그리고 다른 연구 과제에 접근할 수가 있어서 연구자의 연구 결과를 극대화할 수 있음
. 이전 연구의 고품질 디지털 연구 결과를 인용함으로써 다른 연구자들이 쉽게 발견하도록 할 수 있음
. 연구 과제와 연결된 요약 정보를 통해 데이터를 내용적으로 이해할 수 있음
. 다학제에 걸쳐서 연구 데이터의 이용과 중요성을 추적할 수 있음
. 학술적인 연구 성과물들을 완전히 이해할 수 있음
. 표준 인용 포맷을 통해 데이터 연구에 적절하게 기여할 수 있음
<도서관 사서들의 이점>
. 연구자들에게 하나의 자원을 제공함으로써 학술적 연구 성과물을 완전히 이해하도록 할 수 있음
. 연구자들이 자신들의 연구를 발전, 가속시키기 위해 관련 데이터에 쉽게 접근하도록 할 수 있음
. 연구자가 속한 기관에서 출판 형태의 내용을 초과하는 연구 성과물의 전체 파급 효과를 공개할 수 있음
. 데이터 인용 사례가 표준화됨에 따라 연구자들에게 자신들의 연구 기여도를 측정할 수 있는 일관되고 가시적인 도구를 제공할 수 있음
. 연구자들의 연구를 공개할 수 있는 리포지토리를 지정함으로써 연구자들이 데이터 관리계획을 만들 수 있도록 도와줄 수 있음
<연구비 지원기관들의 이점>
. 연구비를 지원한 중요한 연구들을 보다 잘 공개할 수 있음
. 연구비 지원기관이 지원하여 생성된 데이터를 다른 연구자들이 발견할 수 있음
. 현재 이용 가능한 데이터의 검색이 가능하고 동일 연구의 중복 지원을 방지할 수 있으며 데이터의 재해석이 가능함
<학술 출판물을 가진 기관>
. 기관이 가진 데이터 자산의 재이용 현황을 추적할 수 있음
. 기관이 가진 데이터 자산을 더욱 외부에 노출시킬 수가 있음
. 데이터 관리 시설에 더욱 투자할 수 있음
. 국제적으로 인정된 인용 지수 서비스를 이용할 수 있음

15 데이터 인용 가이드라인과 도구

- 일반적인 가이드나 매뉴얼은 데이터를 자원 타입에 포함시키지 않으나, 어떤 저널이나 데이터센터 혹은 리파지토리, 그리고 어떤 학회들은 데이터를 인용하는 방법에 대한 전문화된 지침서를 제공하고 있음[109]
- Dryad는 “데이터 인용에 관한 표준과 제안”이라는 리스트를 제공하고 있으며

109) https://library.uoregon.edu/datamanagement/citingdata.html, (2016. 9.)

그 외에도 데이터 셋 인용을 위한 American Geophysical Union (AGU)의 저자 가이드라인[110]과 같은 지침서와 예제들이 있음

- 대부분의 서지 정보 관리 소프트웨어 프로그램은 데이터 셋을 인용하기 위한 템플릿을 제공하지 않으나, 아래와 같은 인용 도구들은 데이터 셋에 인용을 저장하는데 사용될 수 있음[111]
  . Endnote
  . Zotero : The Next-Generation Research Tool
  . Mendeley

110) http://publications.agu.org/author-resource-center/publication-policies/data-policy/data-policy-faq/, (2016. 9.)
111) https://library.uoregon.edu/datamanagement/citingdata.html, (2016. 9.)

V 결론

- 최근 오픈 사이언스 운동과 함께 정부 부처, 연구비 지원기관 등에서는 공적 기금으로 연구를 수행하는 경우 연구자가 연구 과정에서 생성된 각종 연구 데이터를 관리하고 공개하도록 의무화하려는 움직임이 있음
- 학술 분야의 연구 논문을 심사하고 출판할 때에도 연구의 근거가 되는 데이터를 함께 제출하도록 요구하는 학회나 출판사들이 증가하고 있음
- 데이터에 DOI와 같은 식별자를 부여하여 데이터 리파지토리를 통해 출판하고 연구자들이 이를 인용할 경우 데이터는 학술 커뮤니티에서 중요한 역할을 담당할 수 있으며 데이터의 출판과 인용은 관련 이해당사자들에게 많은 이익을 가져다 줄 수 있음
- 데이터 출판의 중요한 특성으로서는 데이터에 대한 ① 활용성, ② 문서화, ③ 인용, ④ 검증의 4가지 요소를 들 수 있으며, 데이터 출판 순서는 ① 출판 계획 수립, ② 출판 신청, ③ 데이터 및 미디어 파일 제출, ④ 품질 심사와 편집, 표준 기입, ⑤ 출판, 리파지토리 저장과 공개의 순서로 데이터를 출판함
- 데이터의 인용은 저널 논문에 참고문헌으로 서지정보를 인용하는 것처럼 데이터를 인용하는 것을 의미하며 전통적인 학술논문의 출판과 인용의 절차를 따라야 데이터가 학술 커뮤니티에서 연구성과물로서 인정될 수 있으며, 또한 데이터는 학술 논문의 근거가 되는 것이므로 논문과 데이터는 영구 식별자로 서로 연계되어야 함
- 데이터의 중요한 인용 원리로서는 ① 중요성, ② 신뢰와 기여, ③ 증거, ④ 고유 식별, ⑤ 접근, ⑥ 영속성, ⑦ 특정과 검증, ⑧ 상호 운용성과 유연성을 들 수 있음
- 데이터의 인용을 활성화하기 위해 주제별 또는 기관별 리파지토리나 데이터센터에서 표준적인 인용 방법과 인용 요소들을 발표하고 있으며 중요한 인용 요소들은 저자, 출판일시, 제목, 버전, 자원 유형, 출판자, 식별자, 소재지 등이 있음
- 이 밖에도 데이터의 인용에서 고려해야 할 사항으로서 데이터의 입도, 비공개 데이터의 인용, 샘플, 물질과 같은 물리적 데이터의 인용, 시간의 경과에 따라 변화하는 데이터의 인용 등이 있음
- 앞으로 과학연구의 공개, 개방화가 더욱 추진되면 더욱 많은 연구데이터의 공유 활동이 일어날 것으로 예상되며 분야별 또는 유형별로 국제적인 데이터 리파지토리들이 출현하여 학술 논문의 근거가 되는 데이터 저장소로서의 역할을 수행할 것으로 생각됨
- 또한 데이터에 DOI 식별자가 부여되어 학술 논문과 데이터의 상호 연계와 인용을 통한 인용지수 측정, 데이터 간의 상관관계를 기술함으로써 유사 데이터의 검색 등 다양한 데이터 서비스 개발이 진행될 것으로 생각됨
- 국제적으로는 데이터 포털 서비스, 데이터 리파지토리 레지스트리 서비스 등 데이터의 상호 운용을 위한 표준 제정과 학제간 데이터의 통합이 진행되고 다학제의 데이터 분석을 통한 새로운 지식이 다양한 학문 분야에서 창출될 것으로 전망됨