과학 데이터의 출판과 인용에 관한 연구

A Study on Publishing and Citing the Scientific Data

이 상 호, 이 유 선*
한국과학기술정보연구원, 과학기술연합대학원대학교*

Sang-Ho Lee, Yu-Seon Lee*
Korea Institute of Science and Technology Information, University of Science and Technology*

요약

과학적 논문의 기초가 되는 과학 데이터들은 국가 차원에서 체계적으로 관리가 되지 않아 원활하게 유통 및 재활용이 되지 않고 있으며 논문 원문과 데이터들이 서로 연계, 통합되지 않아서 과학 데이터의 유통이 더욱 어려워지고 있다. 이 연구에서는 유통 측면에서 과학 논문과 데이터를 비교하고 연계, 통합을 위한 영구식별자의 도입 및 제도적 측면에서 데이터 기탁 및 데이터 리파지토리의 활성화 방안에 대해 논의한다.

Ⅰ. 서론

학술 논문을 통해 출판된 지식은 과학 데이터로부터 만들어진 마지막 단계의 결과물이며, 과학 데이터들은 분석, 합성, 해석되어 이 과정의 성과물로서 일반적으로 과학적 논문으로 발표된다.

과학 저널에는 원시 데이터의 아주 적은 부분만이 실리게 되며, 대부분의 과학 데이터들은 보안이 잘 되어 있는 기관 리파지토리에 보관되기 보다는 주로 연구실 또는 개인 연구자 단위로 파일로서 관리되고 있다. 따라서 대부분의 데이터가 쉽게 유실될 수 있는 상황에 놓여 있는데 2009년 한국의 생물학연구정보센터에서 실시한 설문조사[1]나 호주의 극지데이터 관리 실태조사[2]에 따르면 연구 과정에서 데이터의 관리 미비로 많은 과학 데이터가 연구 현장에서 유실되고 있으며 이로 인해 연구에 많은 지장이 있음을 보고하고 있다.

이와 같이 연구과정에서 취득된 데이터들이 널리 일반에게 공개되지 못하고 연구실 또는 개인 단위로 관리되거나 폐쇄된 시스템에서 제한된 데이터 공유가 이루어질 경우 동료 연구자들은 이러한 데이터에 접근할 수가 없게 되며, 일부 과학 데이터들이 웹을 통해 공개된다고 하더라도 사이트 주소가 변경되거나 연구과제가 종료되면 더 이상의 데이터 관리가 이루어지지 않아서 결국 시간이 경과하면 해당 과학 데이터에의 접근이 사실상 불가능해 지는 경우가 많다.

이와 같이 과학 데이터에 대한 접근이 어려워지면 연구자들은 불필요하게 중복 연구를 수행하거나 또는 연구에 활용된 원시 데이터에 접근할 수가 없어서 연구자들 사이에 논문으로 발표된 연구 결과를 검증하는데 많은 어려움이 있을 수 있고 국제적으로도 데이터의 공유가 이루어지지 않으면 새로운 국제 공동 연구 과제를 도출하거나 공동연구를 진행하는데 큰 방해물로 작용할 수가 있다. 또한 정부는 매년 엄청난 규모의 연구 예산을 이미 존재하는 과학 데이터를 재생산하는데 사용하므로 연구 예산의 낭비를 가져올 수 있고 데이터의 재활용이 이루어지지 않아 연구 효율성이 떨어지는 결과를 초래할 수 있다.

본 연구에서는 과학 데이터의 유통 특성을 과학 논문과 비교해서 그 문제점을 살펴보고 데이터의 유통을 원활히 하기 위해 원문과 데이터의 연계 통합을 위한 영구식별자의 도입, 그리고 데이터 리파지토리의 구축을 통해 데이터 공유 방안에 대해 논의하고자 한다.

Ⅱ. 과학 논문과 데이터의 유통 특성

과학자들이 저술한 저작물(출판물)인 논문, 연구보고서, 학위 논문 등은 도서관 또는 출판사들이 보유하고 관리하며 각자 보유하고 있는 출판물들을 카탈로깅하여 메타데이터로서 공개하고 있으므로 과학자들은 해당 웹사이트를 방문하여 쉽게 검색을 통한 원문 접근이 가능하다. 또한 상호 인용이 가능하며 이는 임팩트팩터(Impact Factor)의 근간이 되고 있다.

이와 달리 과학논문 작성에 근간이 되었던 과학데이터는 대부분 개인 또는 연구실 단위로 관리하고 있거나 보다 체계적 방법으로 데이터센터에서 관리, 보유하고 있는 경우가 많다. 그러나 데이터센터가 보유하고 있는 데이터셋(Datasets)과 이것을 활용한 과학논문을 연결해 줄 수 있는 수단이 존재하지 않고 단지 비공식적이 네트워크에 의해 데이터의 1차 공유가 이루어지고 있으며 이들을 식별하거나 인용할 수 있는 국제적 통용 규범이 존재하고 있지 않다.

1. 문제점

학계에서는 주로 출판을 통해 성과를 인정받는 반면 데이터셋(Datasets)을 공유하는 데에는 많은 시간이 소요되고 충분한 보상이 어렵고 데이터 출판에 대한 연구자의 의무규정 미비와 불명확한 데이터셋의 소유권, 데이터를 저장할 수 있는 리파지토리의 부재가 데이터 공유의 방해요인으로 작용하고 있다.

또한 논문은 적절한 심사를 거쳐 품질을 인정받고 도서관이나 출판사에 영구적으로 보관되고 인용이 될 수 있으나 과학데이터는 출판되지 않거나 개인 연구자의 웹사이트(URL)를 통해서만 공유되므로 dead link 발생 시 데이터의 공유가 어렵다.

2. 원문과 데이터의 연계 통합

대부분의 데이터셋은 커뮤니티 고유의 등록기관이나 URL과 같은 위치정보를 통해서 직접 인용이 되고 있고 앞서 언급한 바와 같이 dead link가 발생하면 해당 데이터는 더 이상 활용할 수 없다. 데이터셋의 위치는 시간에 따라 바뀔 수 있으므로 데이터 장기 보존에 문제가 있을 수 있고 커뮤니티 고유 등록기관들은 분석 서비스의 이질성 때문에 상호운용성의 문제가 발생한다. 그러나 과학데이터에 URL과 같은 단순 위치 정보가 아닌 영구식별자를 적용하여 원문과 데이터를 통합하고 제한없이 인용할 수 있다면 데이터셋을 출판하는데 촉진제로 작용할 수 있다. 또한 이를 통해 해당 논문을 인용하고자 하는 연구자들이 실제 데이터셋을 확인하고 연구결과를 검증할 수 있고 다양한 측면에서 동일한 문제의 공동연구도 가능해지며 다학제적 연구가 가능하게 된다.

3. 데이터에의 영구 식별자 도입

DOI(Digital Object Identifier)나 URN(Uniform Resource Names)과 같은 영구 식별자를 통해 전자자원을 식별하려는 것은 참고문헌들의 장기보존을 위해 잘 알려진 방법이다. 이와 같이 영구 식별자는 디지털 환경에서 지적재산권을 확실히 식별하고 형식에 무관하게 이러한 지적재산권을 운영할 수 있도록 한다. 현재 세계적으로 가장 많이 쓰이고 있는 식별자는 URN, ARK, PURL, DOI가 있으며 이 중 DOI는 과학 논문을 포함하여 디지털 저작물의 이용 및 관리체계로서 응용되고 있다. 이러한 DOI의 데이터에 대한 적용은 독특하고 독립적인 과학적 객체로서 취급될 수 있는 과학데이터가 연구자들에 의해 인용이 가능하다는 것을 보여준다. 그 예로서 독일의 German national library of science and technology(TIB)는 World data center와 협력하여 약 60만여 건이 넘는 데이터셋에 영구 식별자로서 DOI를 등록한 바 있다. 이는 과학출판물의 한 부분으로서 약 1,500개 이상의 데이터 셋이 선정되어 TIB의 온라인 카탈로그와 German Common Library Network(GBV)를 통해 접근이 가능하도록 하였다.

마찬가지로 우리나라도 DOI를 부여하는 전문기관에서 데이터에 대한 식별자도 부여하여 해당 연구자들이데이터를 상호 인용하고 평가할 수 있는 체계를 만들어야 한다.

Ⅲ. 데이터 리파지토리의 구축

1. 과학 데이터의 카달로깅

메타데이터는 정보자원의 속성들을 기술하는 데이터로서 '데이터에 대한 데이터'이다. 정보자원에 대한 메타데이터를 생성시킬 경우 보다 정확하게 검색할 수 있는 장점이 있지만 반대의 경우 데이터는 숫자, 변수, 그림, 문자 등 각 요소의 나열이 될 뿐이다. 개인 연구자가 연구과정에서 취득한 과학데이터를 일반에게 공개하려면 우선 취득한 과학데이터의 속성을 기술할 수 있는 메타데이터 카달로깅 작업을 수행해야한다. 그러나 개인 연구자들은 자신들이 취득한 과학데이터에 대한 메타데이터의 형식을 잘 모르는 경우가 많으며 분야별 표준이 존재하지 않아 작업이 수월하지 않으므로 공개하고자 하는 데이터를 등록할 수 있는 리파지토리를 설계하면서 이에 대한 분야별 속성을 정의하여 적합한 메타데이터 표준을 배포해야 한다.

2. 데이터 리파지토리 구축

데이터 공유에 있어 가장 큰 문제점은 데이터를 등록하거나 저장할 수 있는 저장소가 없다는 것이다. 현재 과학데이터는 개인연구자, 연구실, 연구그룹 단위로 폐쇄적인 관리가 되고 있으며 기관이 보유하고 있는 데이터 또한 검색시스템에서 검색, 조회가 가능하지만 원시데이터 자체는 폐쇄적으로 관리되고 있다. 그러나 연구자 간의 데이터 공유를 위해서는 폐쇄적인 시스템을 벗어나 원시데이터나 이를 가공한 데이터를 모두 공유할 수 있는 분야별 리파지토리를 구축해야 하며 동시에 이를 운용할 수 있도록 연구개발 주체들이 모여 만들어진 데이터센터 또한 같이 고려되어야 한다. 이를 통해 데이터 기탁에 대한 문화를 자연스럽게 형성하고 이를 통해 관련 학문 간의 데이터 공유가 원활히 이루어 질 수 있다.

Ⅳ. 결론

데이터는 과학 발전에 핵심 역할을 수행하며 데이터 기반의 협업 연구가 증가하고 있는 만큼 데이터 공유의 중요성은 커지고 있다. 데이터에 식별자를 도입하여 원문과 통합하고 리파지토리를 구축함으로써 연구자들이 데이터를 기탁 및 출판하고 상호 인용할 수 있는 통로를 만들 수 있다. 그러나 연구자 개인의 의지뿐 아니라 데이터 기탁에 대한 강제성을 지닌 제도적 지원도 함께 고려되어야 할 것이다.

■ 참 고 문 헌 ■

[1] Review of the Australian Antarctic Data Centre, Nov. 2003, pp. 6.
[2] http://data.aad.gov.au/
[3] http://polaris.nipr.ac.jp/~dbase/