연구데이터 관리 및 검색을 위한 스키마 클래스 상속 모델

Schema Class Inheritance Model for Research Data Management and Search

김선태 (Suntae Kim)*

초 록

최근 연구데이터가 국가자산이라는 인식의 확산으로 원시데이터 관리 및 재사용의 필요성이 이슈이다. 본 연구에서는 데이터의 체계적인 관리를 위해, 스키마 클래스를 상속하는 방식의 메타데이터 설계 모델과 상속을 통해 생성된 스키마 객체들을 대상으로 메타데이터 통합 검색 모델을 제안하였다. 스키마 클래스를 상속한 스키마 객체가 데이터 컬렉션에 1대1의 관계를 갖도록 데이터 아키텍처를 설계하였으며, 제안된 모델의 검증을 위해서 가상 스키마 클래스 및 객체가 시스템적으로 구현 가능함을 증명하였다. 본 연구에서 제안하는 스키마 클래스 상속 및 통합검색 모델은 일반적으로 사용되는 '하향식 계층 모델'의 단점을 극복하는 모델로서, 정부 기관에서 생산되는 데이터를 독립적으로 관리하는데 활용될 수 있다고 사료된다.

ABSTRACT

The necessity of the raw data management and reuse is issued by diffusion of the recognition that research data is a national asset. In this paper, a metadata design model by schema class inheritance and a metadata integrated search model by schema objects are suggested for a structural management of the data. A data architecture in which an schema object has an 1 : 1 relation to the data collection was designed. A suggested model was testified by creation of a virtual schema class and objects which inherit the schema class. It showed the possibility of implement systematically. A suggested model can be used to manage the data which are produced by government agencies because schema inheritance and integrated search model present way to overcome the weak points of the 'Top-dow Hierarchy model' which is being used to design the metadata schema.

키워드: 연구데이터, 스키마 클래스 상속, 연구데이터 관리, 연구데이터 검색
research data, schema inheritance, research data management, research data search

* 한국과학기술정보연구원 과학기술빅데이터연구실 선임연구원(stkim@kisti.re.kr)

■ 논문접수일자 : 2014년 5월 17일 ■ 최초심사일자 : 2014년 5월 28일 ■ 게재확정일자 : 2014년 6월 20일
■ 정보관리학회지, 31(2), 41-56, 2014. 〔http://dx.doi.org/10.3743/KOSIM.2014.31.2.041〕

1. 서 론

대량으로 생산되는 데이터는 데이터가 연구의 중심도구로 사용되는 제 4세대 연구패러다임, data-intensive science를 만들고 있다. 과학기술의 발전으로 연구 환경이 선진화되고 다양한 융・복합 연구가 확산되면서, 대량으로 생산되는 데이터의 보존과 데이터의 재사용을 위한 환경 구축이 요구된다. 이와 관련하여 관측, 관찰, 실험, 조사, 분석, 시뮬레이션 등의 연구활동을 통해서 생산되는 원시데이터(raw data)도 국가의 자산으로서 관리되고 공유되어야 한다는 인식이 일반화 되고 있다. 이를 위해서, 해외 선진국의 움직임이 활발하다. 미국과 독일은 데이터의 글로벌 유통체제를 선도해 나가고 있으며, 호주와 중국은 자국에서 생산되는 데이터를 체계적으로 관리하기 위한 국가적 인프라를 구축해 나가고 있다. 핀란드는 연구자들을 위한 연구데이터를 국가차원에서 공동구매하고 있으며, 캐나다는 자국의 연구데이터 관리체제를 구축하고 있다.

이와 같이, 연구데이터 관리의 중요성이 이슈화 되면서 정책연구가 활발하다. 영국의 연구비 지원기관, 9개 기관[1] 모두 정책조항으로서 연구데이터와 연구기록물로의 접근환경을 제시하기 위한 항목으로 데이터 접근・공유 및 장기 보존 항목을 포함하고 있다(Jones, 2012). 미국의 연구비 지원 주요 기관 10개의 22개 데이터 정책을 대상으로 데이터 정책의 요구사항과 도서관에 대한 시사점을 도출한 Dietrich (2012)는 에이전시 차원의 정책과 기관 수준의 데이터 관리 정책이 필요함을 강조하였다.

한편, 연구자 커뮤니티를 대상으로 데이터에 대한 인식 조사 또한 활발하다. Science(2011) 출판사에서 수행한 설문조사에 의하면, 응답자 1,700명 중 44%가 타 연구자의 데이터를 연구에 사용하고 있다. Tenopir(2011)는 관측분야 1,329명의 연구자를 대상으로 데이터 공유에 관한 설문조사에서 응답자 중 76% 연구자는 제공받은 데이터로 새로운 데이터세트를 생성할 수 있다고 응답하였다. Kuipers(2009)의 연구자 1,389명을 대상으로 한 설문조사에서 연구자 응답자 중 91%는 데이터의 재분석 가능성을 데이터 보존의 핵심 동인으로 생각하였다. 이는 데이터의 재사용이 가능하도록 데이터 보존이 이루어져야 한다는 것을 의미한다. 데이터의 재사용을 위해서는 데이터 생산과 관련된 다양한 맥락(context) 정보가 함께 관리되어야 한다.

이상과 같이, 국가 자산인 연구데이터의 체계적인 관리와 데이터의 공유 및 확산을 지원하기 위해서는 다양한 요소가 필요하다. 무엇보다도 법과 제도가 뒷받침이 되어야 하며, 시스템적인 인프라 구축과 데이터를 압축, 전송하는 기술적 측면의 노력이 필요하다. 또한 연구데이터의 기술(description) 및 검색, 재사용을 위해서는 메타데이터 스키마의 재사용 모델이 필요하다.

본 연구는 연구데이터를 위한 메타데이터 스

1) AHRC(Arts and Humanities Research Council), BBSRC(Biotechnology & Biological Sciences Research Council), RCUK(Cancer Research UK), EPSRC(Engineering and Physical Sciences Research Council), ESRC(Economic and Social Research Council), MRC(Medical Research Council), NERC(Natural Environment Research Council), STFC(Science and Technology Facilities Council), Wellcome Trust.

키마 클래스 상속 모델을 설계, 구현하고 이질적인 연구데이터의 통합검색 방법을 제안하는데 그 목적이 있다. 이를 위하여 본 연구에서는 1) 연구데이터의 특징을 조사하고, 2) 메타데이터를 설계하는데 일반적으로 사용되고 있는 '하향식 계층 모델'의 단점을 조사하였다. 3) 메타데이터를 설계하고 통합검색을 위한 모델로 스키마 클래스 상속 방식의 메타데이터 설계 모델 및 스키마 객체를 이용한 통합검색 방안을 제안하였으며, 4) 제안하는 모델이 시스템적으로 구현 가능함을 증명하였다.

2. 관련 연구

스키마 클래스 상속 방식을 적용한 메타데이터 설계에 관한 연구사례는 발견하기 힘들다. 따라서 본 연구에서는 연구데이터의 관리와 공유를 위한, DataCite 컨소시엄에서 제정한 메타데이터를 조사하였다. 또한 리포지터리 시스템으로 전 세계에서 가장 많이 활용되는 DSpace 시스템의 메타데이터 모델을 조사하였으며, 일부 관련 연구를 조사하였다.

학술 논문에 디지털객체식별자(DOI)를 부여했듯이 데이터에도 DOI를 부여해서 데이터 유통체제를 선도해 나가는 DataCite가 최근에 메타데이터 표준을 개정 발표하였다. Starr, Ashton, Barton, Elliott, Jacquemot-Perbal, Karjalainen, McAvoy, Newbold, Peters, Smaele, Schleinstein, Zenk-Möltgen, Ziedorn, F(2013)는 2011년 3월에 처음 발표된 DataCite 메타데이터 스키마를 재개정하여 버전3을 발표하였다. DataCite 스키마의 최상위 요소는 18개로서, 'Identifier' 요소를 포함한 필수속성 5개와 'Subject' 요소를 포함한 권고속성 7개, 'Language' 요소를 포함한 선택속성 6개로 구성되어 있다. DataCite 메타데이터는 연구데이터를 기술하기 위한 메타데이터 표준을 제시하고 있다. 특히, 최상위 요소로서 Identifier, Creator, Title, Publisher, PublicationYear 요소들을 데이터 출판을 위한 필수 요소로 지정하였다. 단위 기관에서는 독자적인 메타데이터 스키마로 데이터를 관리하고 DataCite에 메타데이터를 공개하고자 할 경우에 해당 스키마로 매핑 된 데이터세트를 전달해 주면 된다. DataCite는 Juan-le, Song-cai, Chuan-jie(2005)가 사용한 '하향식 계층 모델'이라 구분될 수 있다.

대표적인 리포지터리 등록서비스인 OpenDOAR과 ROAR에서 제공되는 통계에 의하면, 연구데이터 관리 및 서비스에 사용되는 솔루션으로 DSpace가 가장 큰 점유율을 보이고 있다. OpenDOAR과 ROAR에 등록된 DSpace 사용기관은 2014년 3월 25일 현재, 각각 1,087개, 1,446개로 집계된다. DSpace가 가장 많이 사용되는 이유는 설치 및 사용이 용이하기 때문이다. 하지만 DSpace의 강점이자 약점은 'Dublin Core(이하, DC)'를 기본으로 한, 상속이 불가능한 스키마 사용에 있다. 이는 한 가지 유형의 자원을 관리하거나, 연구레코드 중심의 자원을 관리함에 있어서는 큰 문제가 없었으나, 다양한 원시데이터를 기술하기 위한 스키마로 사용하는데 있어서는 한계를 가지고 있다. 즉, 모든 데이터에 동일한 메타데이터 요소와 동일한 표시상수를 사용해야 하는 단점을 가지고 있다. DSpace에서 제공하는 메타데이터 설계 모델도 DC를 기반으로 하는 '하향식 계층모델(Top-down hierarchy

model)'이라 할 수 있다.

Vempati, D. U., Chung, C., Mader, C., Koleti, A., Datar, N., Vidović, D., Wrobel, D., Erickson, S., Muhlich, L. J., Berriz, G., Benes, H. C., Subramanian, A., Pillal, A., Shamu, E. C., Schürer, C. S.(2014)는 LINCS[2]에서 생산되는 다양한 데이터를 관리하고 서비스하기 위한 통합 메타데이터를 설계하였다. 최소한의 메타데이터를 설계하고, 통제어와 온톨로지를 제시하여 사용을 권장하였다. Juan-le, Song-cai, Chuan-jie(2005)는 지구과학 분야의 데이터를 공유하기 위한 메타데이터를 설계하기 위해서, '하향식 계층 모델'을 사용하여 핵심과 스키마, 규격 메타데이터로 구성된 메타데이터를 설계하였다. 이들은 많은 어플리케이션 응용파일이 핵심메타데이터를 확장하여 사용할 것을 기대하였다.

한편, 국내에서 김선태(2012)는 '하이브리드형 메타데이터 요소도출 방법'을 통해서, 해양관측 분야의 과학데이터를 검색하고 발견하기 위한 메타데이터 요소를 도출하는 것과 도출된 메타데이터 요소를 활용하여 확장성 있는 데이터베이스 객체모델을 제안하였다. 그 외 국내의 지구관측자료 공유를 위한 메타데이터 연구(이혜영, 2007)와 생물다양성 데이터를 교환하기 위한 메타데이터 스키마 설계(안부영, 2005), 실험, 측정, 시뮬레이션 데이터세트에 한정된 '과학기술 데이터세트 메타데이터 표준 개발'(이상태, 2006), 해양조사데이터를 포함한 각종 문헌정보를 통합 관리할 수 있는 메타데이터 설계(한종엽, 2004) 등의 연구가 진행되었다.

메타데이터 스키마에 관한 연구는 매우 활발하게 진행되고 있다. 2014년 5월 현재, 구글에서 'metadata schema' 키워드를 이용해 검색한 결과, 128,000건 이상이 검색되는 정도이다. 하지만 'metadata schema inheritance' 혹은 'schema inheritance' 키워드로 검색을 하게 되면, 본 연구에서 제안하는 메타데이터 스키마 클래스 및 스키마 객체 생성과 관련된 유사 연구를 찾기 힘들다. 이는 본 연구가 개척 분야의 연구임을 의미한다. 따라서 본 연구는 메타데이터를 설계하는 새로운 접근 방식을 제시하는 의미 있는 연구로 판단된다.

3. 연구데이터의 특징

연구데이터는 과학데이터를 포괄하는 개념이다. 과학데이터는 논문, 특허, 보고서 등과 같은 연구레코드를 제외한 원시데이터 중심의 데이터를 의미한다. 김선태(2012)의 정의에 따르면, 과학데이터란 연구자의 연구 활동 과정 중 생성되는 다양한 유형의 사실적 기록을 의미한다. 즉, 연구 활동을 통하여 생산된 연구 활동의 기록물로서 관측, 감시, 조사, 실험, 분석, 계산 등의 과정을 통하여 생산된 문자, 이미지, 오디오, 동영상 등의 아날로그 및 디지털 형식을 포괄하는 데이터이다. 본 연구에서는 연구데이터와 과학데이터를 동일한 개념으로 접근한다. 다만, 논문, 특허, 보고서 등과 같은 연구레코드는 원칙적으로 연구데이터 범주에 포함되지만, 본 연구에서는 연구레코드가 제외된 개념으로

2) LINCS는 The National Institutes of Health Library of Integrated Network-based Cellular Signatures를 의미한다.

그 의미를 한정한다.

연구데이터는 형태에 따른 분류로 디지털 데이터와 아날로그 데이터로 크게 나눌 수 있다. 디지털 데이터는 관측・측정 장비, 분석 장비 등에 의해 생산되어 유형적인 형태로 존재하지 않고 디지털 저장장치에 파일형태로 저장된 데이터이다. 아날로그 데이터는 연구자의 채집 및 수집 활동 또는 채집 장비, 시추기 등을 통하여 획득된 실질적인 형태를 지니고 있는 시료, 표본, 생물 등의 데이터이다(김선태, 2012). 과학기술의 발달로 디지털 연구데이터의 생성속도도 가속화 되었다. 이로 인해 과학기술분야의 빅 데이터 문제가 제기되는 것이다. 한편 연구데이터는 데이터의 생성부터 처리과정에서 생산되는 절차를 기준으로 분류될 수 있다. 최초 데이터의 획득 이후, 추가적인 처리 및 변경이 없는 원시 데이터와 데이터에 대해 처리, 실험 또는 분석 등의 과정을 거쳐 2차로 생산되는 중간데이터(processed data), 최종 결과데이터(result data)로 분류될 수 있다. 하지만 연구의 과정에서 연구 소스로 사용되는 중간데이터는 해당 연구의 기준에서는 원시데이터가 될 수 있으므로 그 기준이 모호하다.

연구데이터는 동일한 연구 대상에 대해서도 연구 분야에 따라 다양한 주제 분야로 구분될 수 있으며, 하나의 주제 분야에서도 다양한 형태의 데이터가 생산된다. 극지연구의 예를 들면, 극지라는 연구 대상을 갖고서 기상데이터, 생물자원데이터, 위성영상데이터 등 다양한 데이터를 수집・생산하고 있다. 이는 극지를 연구하는 극지연구소 하나의 기관에서 다양한 유형의 데이터를 체계적으로 관리해야 함을 의미한다. 한편, 과학기술의 발달로 생산되는 연구데이터 볼륨이 매우 크다. 쇄빙연구선 '아라온'호에 장착된 주요 연구 장비에서 생산되는 디지털 원시데이터의 경우, 년 26테라바이트가 생산되고 있다(한국과학기술정보연구원, 2011). 2013년 현재, CERN의 강입자충돌기(Large Hadron Collider, LHC)는 1년에 780테라바이트 가량의 데이터를 생산해 내고 있으며, 슬로언 디지털 스카이 서베이(Sloan Digital Sky Survey) 프로젝트는 최근 60테라바이트 데이터를 공개했다(Wiki, 2014b). 그리고 2013년 3월 14일, CERN에서 힉스 보손의 발견을 공식으로 밝히면서 입자는 실재하다는 것이 증명 되었다. 이렇듯 데이터 중심 연구 분야에서는 대량의 데이터를 생산해 내고 있으며, 대용량 데이터를 기반으로 데이터 중심 연구를 통해 의미있는 연구결과를 도출하고 있다. 하지만, 연구자들은 연구데이터를 활용하는데 있어, 원하는 데이터를 획득하는데 어려움이 있으며, 데이터를 획득하더라도 데이터의 사용방법 등이 체계적으로 기술되어 있지 않기 때문에 그 활용이 매우 어려운 실정이다.

데이터 중심 연구를 지원하기 위해서는 도서관이 기존에 서비스하던 논문, 특허, 보고서, 단행본 같은 문헌중심의 연구레코드 뿐 아니라 보다 확장된 연구데이터에 대한 고민이 필요하다. PRC(2010)는 데이터세트와 데이터모델, 알고리즘, 프로그램 콘텐트가 연구자들이 필요로 하지만 접근하기는 어렵다는 것을 주장하였다. PRC는 18,000개 저널의 논문 투고자 51,000명을 대상으로 설문조사를 수행하였으며, 7.5%의 응답률로 3,823명이 응답을 하였다. 저널과 단행본, 참고도서, 컨퍼런스 프로시딩에 이어 데이터세트와 데이터모델, 알고리즘, 프로그램 콘

텐트가 중요하다고 응답하였다. 이는 전통적인 문헌형식의 콘텐트인 연구레코드에 이어 연구데이터에 대한 연구자들의 활용 요구를 나타낸다고 판단 할 수 있다. 한편, 연구데이터로의 접근율은 모든 조사대상 콘텐트 중에 연구데이터가 38%로 가장 낮았다. 원시데이터로의 접근율을 높이기 위해서는 연구데이터의 특징을 감안한 스키마가 필요하다. 원시데이터 자체는 관측, 관찰, 실험, 조사, 분석, 시뮬레이션 등을 통해 생산되기 때문에 콘텐트 자체가 무의미한 이진수, 0과 1의 나열로 구성되는 경우가 많다. 따라서 데이터의 재사용성을 높이기 위해서는 데이터에 대한 상황정보를 메타데이터로 구축해야 한다. 이러한 연구데이터의 특징을 감안하여, 김선태와 이태영(2011)은 고성능 컴퓨터와 네트워크, 자동화 센서, 고성능 관측 및 실험장치가 대량의 연구데이터를 쏟아내고 있기 때문에 연구의 패러다임 변화가 가속화 되고 있다고 판단하였으며, 생산되는 연구데이터를 사장시키지 않고 효과적으로 이용하는 방법이 모색되어야 한다고 주장하였다.

이상과 같이, 다양한 연구 활동을 통해 생산되는 연구데이터의 특징은 크게 5가지로 구분될 수 있다. 연구데이터는 1) 데이터의 생성 방법과 데이터의 처리 절차, 데이터를 생산하는 연구 분야에 따라서 그 형태가 매우 다양하며, 2) 과학기술의 발달로 디지털 연구데이터가 빠르게 생산되고 있다. 3) 생산되는 데이터의 크기가 크며, 4) 데이터 중심 연구에서 대용량 데이터가 새로운 가치를 추출하는데 소스로 사용된다. 하지만 데이터 관리가 쉽지 않기 때문에 5) 데이터 검색 및 재활용이 쉽지 않다.

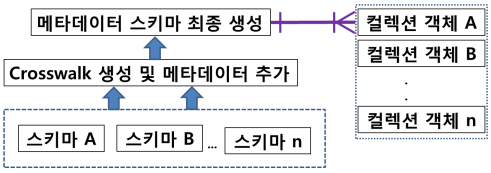
4. 연구데이터를 위한 메타데이터

4.1 하향식 계층모델(Top-down Hierarchy Model) 방식 메타데이터 설계 모델

기관에서 생산되는 데이터의 관리를 위해서 다양한 메타데이터 표준이 생성되고 갱신되어진다. 메타데이터 표준은 데이터를 설명하는 항목들의 구성과 작성원칙이 결합된 스키마로 발표되어진다. 스키마의 작성 목표는 커뮤니티의 요구를 담아서 특정 자료유형을 설명하기 위해 개발되기도 하며, 주제별로 데이터를 관리하기 위해서 개발되기도 한다. 또한 특정 프로젝트의 요구를 수용하기 위해 개발되기도 한다. 이러한 메타데이터의 표준화에 대한 노력은 오래전부터 진행되어오고 있다. DC 이전에 발표된 실질적인 메타데이터 표준으로는 TEI Header가 유일하며, 문헌정보학 분야에서 주장되는 최초의 메타데이터 표준은 1995년에 개발된 DC이다(남태우, 이승민, 2010). 하지만, 지구관측데이터의 메타데이터 표준인 DIF의 경우도 TEI Header와 같이 1987년부터 시작되었으며, 해당 분야에서 20년 이상 지속적인 메타데이터 갱신 작업이 이루어지고 있다. 1990년대 다양한 유형의 메타데이터 표준이 발표되면서 2001년에는 미국의 디지털도서관재단(Digital Library Foundation) 지원 하에 메타데이터를 인코딩하고 전송하는 표준(METS)까지 개발이 되었다. 이후 2000년 중반 이후부터 데이터에 대한 메타데이터 개발이 활발하게 진행되고 있다. 최근에는 2011년 1월에 국제컨소시엄 DataCite에서

연구데이터용 메타데이터 스킴을 발표하였다. 이는 데이터를 기술하기 위한 응용프로파일을 작성할 때 매우 유용하게 사용될 수 있다(김선태, 2012).

이러한 메타데이터를 설계하는 방식으로 1) 자체적인 메타데이터 설계 방식, 2) 이미 공표된 메타데이터 표준을 사용하는 방식, 3) 관련 메타데이터의 요소 의미를 매핑해서 Crosswalk를 작성하고, 메타데이터 요소의 선택과 추가를 통한 설계 방식이 존재한다.



〈그림 1〉 Top-down Hierarchy Model 방식의 메타데이터 설계 모델

〈그림 1〉은 Crosswalk 방식의 메타데이터 설계 모델을 보여준다. 위 방식 중 3)번 방식의 절차를 표현한 것이다. 기관에서 관리할 대상 콘텐트와 관련된 스키마들을 수집한 후, Crosswalk를 작성하고 메타데이터 요소의 선택과 메타데이터의 추가 작업을 거쳐, 최종적인 메타데이터 스키마를 작성한다. 이후, 컬렉션에 적용하여 데이터를 관리하고 서비스 한다.

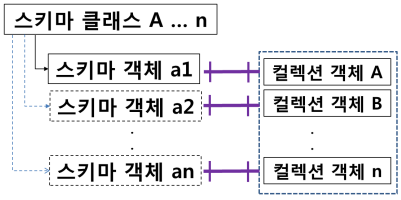
이 상의 모든 방식은 스키마의 재사용성에 문제가 발생한다. 생성한 스키마를 데이터를 관리하는 그룹(이하, 컬렉션)에 일괄 적용하여 사용하거나, 컬렉션 별로 스키마를 재설계해야 하는 작업이 불가피 하다. 또한 스키마의 변경이 일괄적으로 적용되지 못하기 때문에 스키마별로 변경사항을 다시 설정해야 한다. 이는 전 세계에서 리포지터리 시스템으로 가장 많이 사용되는 DSpace[3]에서 채택하고 있는 스키마 모델과 동일하다. DSpace 스키마 모델의 단점은 하나의 스키마로 모든 컬렉션의 데이터를 기술해야하는데 있다. 따라서 연구레코드를 기술하는데 있어서도 문제가 발생한다. 예를 들어, 논문과 특허, 보고서 등을 별도의 컬렉션으로 관리한다고 가정 했을 때, 모든 컬렉션이 모두 동일한 스키마를 사용해야 한다. 이는 이질적인 데이터에 동일한 표시상수를 사용함으로써 데이터 관리자 및 이용자에게 데이터 기술시 혼란을 준다.

4.2 스키마 클래스 상속 방식 메타데이터 설계 모델

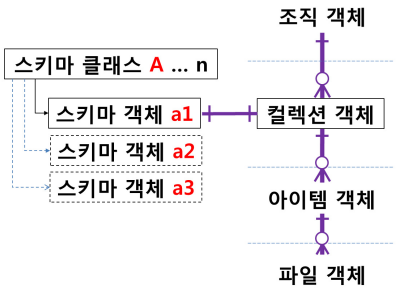
본 절에서는 기존에 사용되는 스키마 생성 방식의 단점을 극복한 스키마 클래스 상속 방식을 제안한다. 〈그림 2〉는 스키마 클래스 상속 방식의 메타데이터 설계 모델을 보여 준다. 객체지향 프로그래밍 언어에서와 동일하게 스키마 클래스를 생성하여 이를 통해 원하는 스키마 객체를 생성할 수 있다. 즉 '스키마 클래스

3) DSpace는 2002년에 미국의 하버드대학과 HP사가 공동으로 개발한 리포지터리 솔루션으로서 오픈소스로 공개된 소프트웨어이다. 2014년 3월 25일 현재, 리포지터리 등록서비스인 OpenDOAR과 ROAR에 등록된 리포지터리 중 DSpace 소프트웨어를 사용하는 건수가 각 각 1,087건과 1,446건으로 전 세계에서 리포지터리 시스템으로 가장 많이 활용되고 있다. 이는 아시아 주요 3국인 대한민국, 중국, 일본에서도 각 각 146건, 224건으로 동일하다. 'DSpace'를 이어 'EPrints', 'Digital Commons', 'OPUS' 소프트웨어가 리포지터리 소프트웨어 도구로 가장 많이 활용되고 있다.

A'를 생성한 후, 이를 이용하여 '스키마 객체 a1', '스키마 객체 a2'를 생성한다. 생성되는 객체에는 클래스에 선언되어 있는 기존 메타데이터 요소의 활용여부를 선택할 수 있으며, 새로운 메타데이터 요소를 추가할 수 있다. '스키마 클래스 A'에 선언되어있는 메타데이터 요소의 표시상수를 클래스의 객체에서 자유롭게 수정가능하다. 생성된 객체는 컬렉션 객체와 1대 1로 매핑이 되어 사용된다.



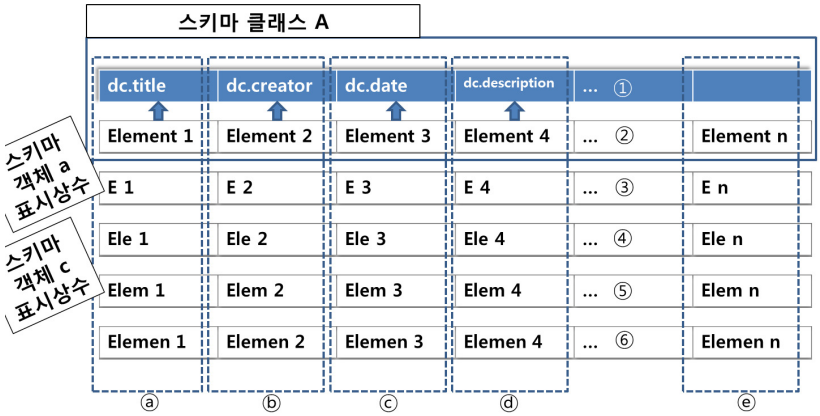
〈그림 2〉 스키마 클래스 상속 방식의 메타데이터 설계 모델



〈그림 3〉 스키마 클래스 상속모델과 데이터 아키텍처의 관계

〈그림 3〉은 스키마 클래스 상속 모델과 데이터 아키텍처의 관계를 나타낸다. 스키마 클래스 상속 모델은 스키마 클래스를 통해서 복수의 스키마 객체가 생성되는 모델이다. 스키마 객체는 컬렉션 객체와 1대 1의 관계를 갖는다. 스키마 객체는 컬렉션이 생성되는 시점에 스키마 클래스를 지정하는 행위로 시작된다. 따라서 컬렉션을 생성하기에 앞서 스키마 클래스가 설계되어야 한다. 데이터 아키텍처는 '조직 객체', '컬렉션 객체', '아이템 객체', '파일 객체'로 구성되며, 계층적 구조로 구성하였다. 이는 전 세계에서 리포지터리 솔루션으로 가장 많이 사용되는 DSpace의 데이터 아키텍처와 동일하다. DSpace에서 사용하는 데이터 아키텍처는 Community, Collection, Item, Bundle로 구성되어(Phillips & Koenig, 2008) 있다. 조직은 하나 이상의 컬렉션을 가질 수 있으며, 하나의 컬렉션은 0개 이상의 아이템을 가질 수 있다. 하나의 아이템은 0개 이상의 파일을 가질 수 있으며, 컬렉션은 1개의 스키마를 가질 수 있도록 설계하였다. 컬렉션 객체는 스키마 객체와 1대 1로 매핑 되므로 컬렉션별로 고유의 표시상수를 가질 수 있다. 또한 스키마 클래스에서 선언되어 있는 메타데이터 요소를 선택적으로 사용 사용할 수 있으며, 추가적인 메타데이터 요소의 추가도 가능하다.

〈그림 4〉는 스키마 클래스 상속 모델을 통한 통합 검색 모델을 나타낸다. ②번은 '스키마 클래스 A'의 요소들을 개념적으로 나타낸다. ①번은 '스키마 클래스 A'에서 설계된 요소들과 1대 1로 매핑 되는 색인 필드를 나타낸다. 색인 필드는 DC의 핵심요소로 지정된다. ③번, ④번, ⑤번, ⑥번은 스키마 객체를 생성할 때 변경된 표시상수를 의미한다. 스키마 객체에서 사용하는 표시상수가 어떠한 형태를 갖더라도 스키마 클래스에서 지정된 색인 필드의 의미로 색인이 생성된다. ⓐ항목의 요소들은 모두 DC의 title 의미로, ⓑ항목의 요소들은 모두 DC의 creator



〈그림 4〉 스키마 클래스 상속 모델을 통한 통합 검색 모델

〈표 1〉 하향식 계층 모델과 스키마 클래스 상속 모델의 비교

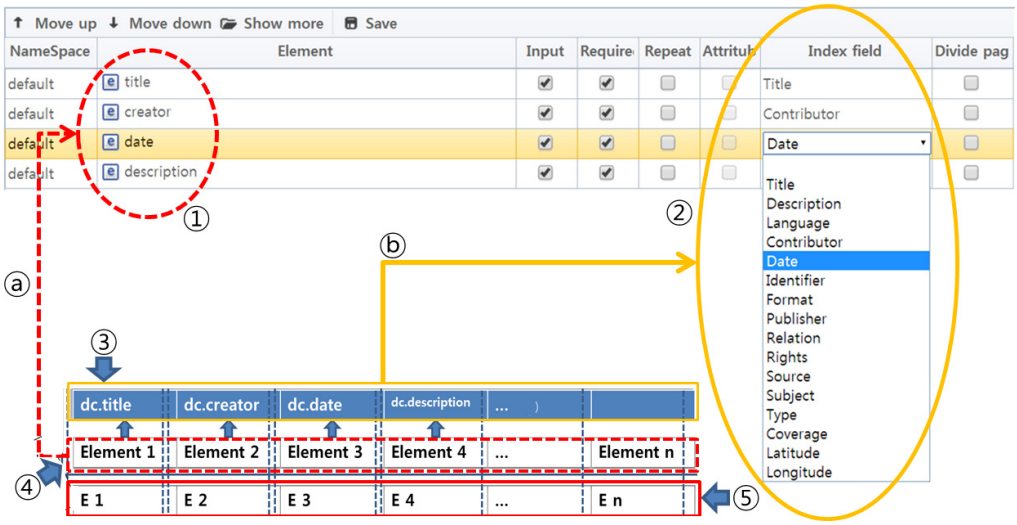
구분	하향식 계층 모델	스키마 클래스 상속 모델
스키마 상속	불가능	가능
표시상수 변경	일부 가능	가능
스키마 재사용	일부 가능	가능
통합 검색	일부 가능	가능

의미로, ⓒ항목의 요소들은 모두 DC의 date 의미로, ⓓ항목의 요소들은 모두 DC의 description 의미로 색인이 생성된다. 이러한 방법을 통해서 '하향식 계층 모델' 방식의 단점을 극복하여 표시상수 변경, 스키마 클래스 상속을 통한 재사용, 다 분야 메타데이터의 통합검색이 가능하다. 〈표 1〉은 이상에서 논의된 '하향식 계층 모델'과 '스키마 클래스 상속 모델'의 특징을 정리하여 보여준다.

4.3 스키마 클래스 상속 방식의 메타데이터 설계 모델 구현 검증

본 절에서는 스키마 클래스 상속 방식의 메타데이터 설계 모델을 시스템 적으로 구현하여 검증하였다. 검증 절차는 다음과 같다. 1) 'Schema Class'를 생성한다. 이후 만들어질 컬렉션에서 통합검색이 가능하도록 색인 필드를 지정한다. 2) 가상 조직(Test Organization)을 만든다. 3) 조직의 자원을 그룹화해 관리할 가상의 컬렉션 3개(Collection-A, Collection-B, Collection-C)를 만든다. 4) 컬렉션을 생생하며, 이미 만들어 놓은 스키마 클래스를 사용하여 스키마 객체를 생성한다. 5) 생성된 컬렉션에 데이터를 입력하고 통합검색 결과를 검증한다.

〈그림 5〉는 '스키마 클래스를 생성'하는 화면으로서 요소를 생성하고 및 통합검색을 위한 색인 필드를 지정하는 화면이다. ①번은 생성



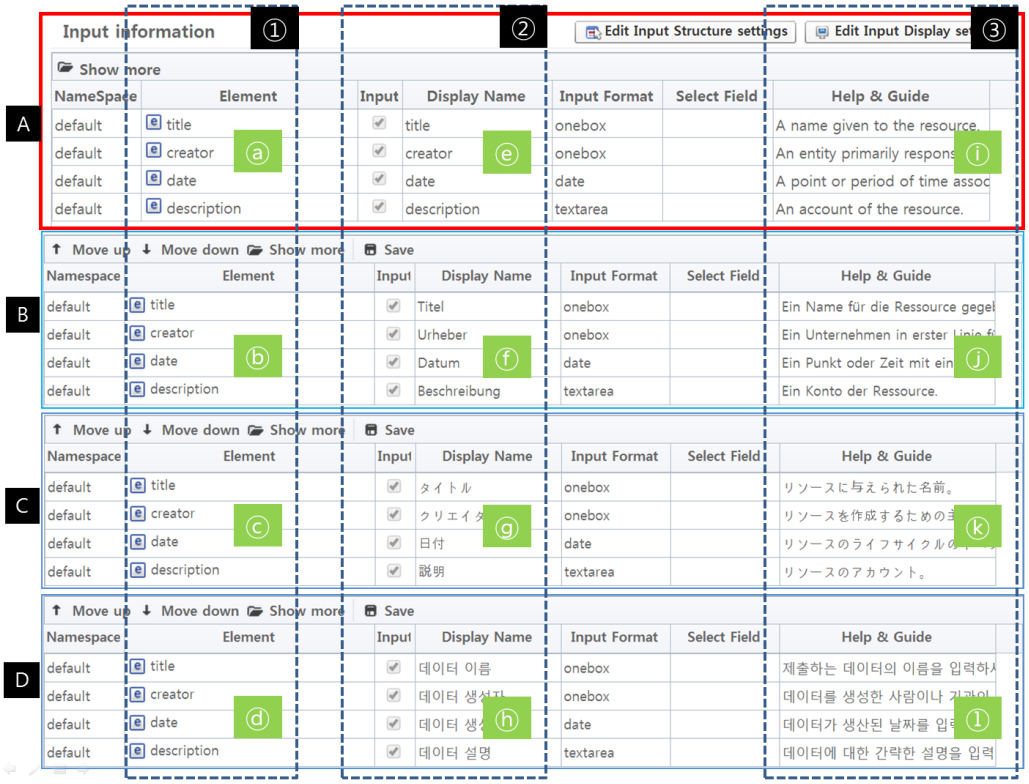
〈그림 5〉 스키마 클래스 생성: 요소 생성 및 통합검색을 위한 색인 필드 지정

되는 스키마 클래스의 요소를 의미하며, 'title'과 'creator', 'date', 'description' 요소가 생성되는 모습을 보여준다. ②번은 생성되는 스키마 요소를 어떤 의미로 색인을 생성할지 지정하는 것이다. DC의 핵심요소를 기준으로 지정가능하다. ③번과 ④번은 〈그림 4〉에서 설명된 스키마 객체의 표시상수와 스키마 클래스의 요소를 나타낸다. ⓐ는 스키마 클래스에서 선언된 요소 개념과 인터페이스와의 관계를 나타낸다. ⓑ는 스키마 클래스에서 선언된 색인 필드의 개념과 인터페이스와의 관계를 나타낸다. ⑤번은 스키마 클래스로 생성한 스키마 객체의 표시상수를 의미한다. 〈그림 6〉은 스키마 클래스 상속을 통해서 생성된 스키마 객체의 생성과 각각의 스키마 객체의 표시상수를 변환한 모습이다. ①번은 스키마 클래스(A)와 스키마 객체(B, C, D)에서 사용하는 메타데이터 요소를 의미하며, ②번은 스키마 클래스와 스키마 객체에서 사용하는 요소의 표시상수를 의미한다. ③번은 스키마 클래스와 스키마 객체에서 사용하는 메타데이터 입력 도움말을 의미한다. 'A'는 스키마 클래스를 의미한다. 'title'과 'creator', 'date', 'description' 요소가 메타데이터 요소로 설계되어 있다. 'B'와 'C', 'D'는 스키마 클래스를 통해 생성한 스키마 객체들을 의미한다. 'B-b', 'C-c', 'D-d'는 모두 'A-a'에서 선언되어 있는 요소를 동일하게 사용하도록 되어 있다. 'A-e(영어)'는 스키마 클래스에 선언되어 있는 표시상수를 의미하며, 'A-f(독일어)', 'A-g(일본어)', 'A-h(한국어)'는 스키마 객체에서 수정된 표시상수를 의미한다. 'A-i(영어)'는 스키마 클래스에 선언되어 있는 메타데이터 입력 도움말을 의미하며, 'A-j(독일어)', 'A-k(일본어)', 'A-l(한국어)'은 스키마 객체에서 수정된 도움말을 의미한다.

본 연구에서 제안하는 모델이 시스템적으로 구현 가능함을 증명하기 위해서, 스키마 클래스 'SchemaClass'를 생성하였다. 해당 스키마

를 사용하는 컬렉션 Collection-A, Collection-B, Collection-C를 생성하여 연구데이터를 한 건씩 입력하였다. 입력된 메타데이터는 모두 스키마 클래스의 요소인 'Title'요소에 'Title'이라는 키워드를 가지고 있도록 샘플 데이터를 입력하였다. 〈그림 7〉은 스키마 클래스 상속을 통해서 생성된 가상의 스키마 객체들을 대상으로 데이터명의 키워드 'Title'을 이용한 통합 검색 화면을 나타낸다. '001' 키워드를 이용해 '데이터이름' 필드를 성공적으로 검색한 결과를 보여준다.

한편, 본 연구에서 제안하는 메타데이터 설계 모델의 적용 주체는 연구데이터를 주체적으로 관리하게 되는 도서관에서 이루어 질 것이다. DataONE 사이버 인프라스트럭처 플랫폼을 디자인하기 위해서 4단계 이용자 참여형 분석을 적용한 Michener(2012)는 DataONE과 이해관계를 갖는 5개 커뮤니티로서 Academia, Community, Government, Non-profit, Private Industry를 도출하였다. Michener는 융합 과학(integrative science)은 데이터연구 중심이며, 정보에 의존적인 특징이 있는 점을 강조하였으며, 도출된 5개 이해관계 그룹 내에서 도서관과 사서가 이미 서비스를 제공하고 있기 때문에, 도서관과 사서를 2단계 이해관계 그룹 중 가장 중요한 그룹으로 정의하였다. 이는 Abrams



〈그림 6〉 스키마 클래스 상속을 통한 스키마 객체 생성과 표시상수 변환

〈그림 7〉 DC를 기반으로 한 스키마 객체의 통합 검색

(2013)의 주장처럼 도서관과 사서가 시간과 공간의 제약 없이 이용자가 원하는 콘텐트를 제공하는 고유의 임무를 수행 하고 있으며, 학술연구 생명주기와 정보 생명주기의 접점에서 솔루션 제공해오고 있는 것을 의미한다. 또한 현재의 도서관과 사서가 정보의 중심에서 서비스를 하고 있으며, 데이터 중심의 융합 과학을 선도하는데 있어 중추적인 역할을 수행 할 수 있음을 의미한다. 한편, ACRL Planning and Review Committee(2012)에서 제시한 미래도서관에 요구되는 10대 트랜드로서 '도서관 가치 증명', '데이터 큐레이션', '디지털 보존'과 Cox와 Pinfield (2013)의 연구에서 도출된 미래도서관 우선순위 중 '리포지터리 관리자 역할'은 도서관과 사서의 미래 모습을 예시한다.

현재의 도서관과 사서는 논문, 보고서, 단행본과 같은 연구레코드 중심의 서비스를 성공적으로 제공해 오고 있다. 이제 PRC(2010)의 연구에서 도출되었듯이 연구자들이 새롭게 요구하는 콘텐트로서 원시데이터 연구데이터를 제

공할 수 있도록 준비해야 한다. Green(2009)은 연구자들이 전체 연구시간 중 데이터 확보에 가장 많은 시간을 투자하고 있다고 주장하였다. 본 연구에서 제안되는 스키마 클래스 상속과 통합검색 모델은 연구자들의 데이터 확보 시간을 단축시킬 수 있다고 판단되기 때문에, 연구데이터를 체계적으로 관리하고 서비스하는데 중요한 컴포넌트로 사용될 수 있다. 기관에서 관리해야 하는 연구데이터를 그룹화 하여 컬렉션을 만들고, 공통으로 사용될 수 있는 메타데이터 항목을 설계하여 스키마를 만들고, 컬렉션별로 해당 스키마를 상속하여 확장・사용할 수 있다. 확장되는 스키마의 표시상수를 수정하여 컬렉션별 데이터 제출 인터페이스를 구성할 수 있으며, DC매핑을 통해서 모든 컬렉션을 대상으로 한 DC기반 통합검색 기능을 구현할 수 있다.

우리나라는 데이터 관리 주체가 여러 부처의 산하기관으로 다원화 되어 있고, 이들간의 상호협력 체제가 제대로 구축되지 않아 잠재적인 효용성에도 불구하고 데이터의 공유 및 재활용이 극히 저조한 수준이다(한국과학기술정보연구원, 2011). 본 연구에서 제안하는 스키마 클래스 상속 및 통합검색 모델은 다 부처에서 생산되는 데이터를 독립적으로 관리하는데 적용될 수 있으며, 국가차원의 데이터 공유・융합 플랫폼을 설계할 때 사용될 수 있다고 판단된다.

5. 결 론

연구환경의 선진화는 데이터관리문제를 야기 시켰지만, 동시에 새로운 연구방법을 도출시켰다. 과거 연구자들은 이론을 수립하고 이를 증명하기 위해 데이터를 생산하였다. 하지만 데이터 중심 연구(data-driven science)에서는 데이터를 기반으로 새로운 이론을 수립하고 테스트 해 보는 것이 가능해 졌다(Mabe, 2009). 연구패러다임이 바뀐 것이다. 이러한 연구패러다임은 eResearch 환경을 만들어내고 있다. 연구의 중심도구로 사용되는 데이터는 eResearch 환경에서는 융합연구의 단초가 될 수 있다. eResearch 개념은 eScience와 사이버인프라스트럭처를 인문학, 사회과학 등의 분야로 확장함으로써 등장하였다. eResearch 분야에서의 연구는 협력이 요구되며, 그리드 컴퓨팅 기술을 사용하고, 연구의 중심도구로 데이터를 사용하는 것이 특징이다(Wikipedia, 2013). 특히, eResearch는 국내에서 2012년부터 본격적으로 논의되고 있는 빅데이터 연구로 더욱 주목 받고 있는 개념이다. eResearch 연구환경에서 연구데이터가 활용되고 공유되기 위해서는 데이터 관리와 서비스가 이를 뒷받침 해주어야 한다. 이를 위해서는 연구데이터를 체계적으로 관리하기 위한 메타데이터 스키마가 필요하다. 본 연구에서는 연구데이터 관리와 통합검색을 위한 메타데이터 설계 시 일반적으로 사용되는 '하향식 계층 모델'을 검토하였다. 또한 이의 단점을 보완한 스키마 클래스 상속 방식(Schema Class Inheritance Model)의 메타데이터 설계방식을 제안하고, 스키마 클래스의 상속을 통한 스키마객체를 DC기반으로 통합 검색하는 방법을 제안하였다. 제안된 모델이 시스템적으로 구현 가능함을 증명하기 위하여 가상의 스키마 클래스와 스키마 객체를 생성하고 샘플 데이터를 입력해, DC 기반의 통합 검색이 가능함을 증명하였다.

한편, 제안된 스키마 클래스 상속 방식의 메

타데이터 설계 모델은 클래스의 재사용은 가능하나 객체의 재사용은 불가능하다. 따라서 동일한 객체를 생성하고자 할 경우, 클래스를 이용한 객체 생성 후 반복적인 작업을 또 수행해야하는 단점이 존재한다. 따라서 향후 연구로서 스키마 객체를 이용한 스키마 재사용에 관한 연구가 필요하다. 또한 정적인 HTML 페이지에 사용되는 태그에 의미를 부여하기 위해서 스키마 클래스 생성 시점에 RDFa를 적용할 수 있는 시스템적인 연구가 필요하다.

참 고 문 헌

김선태, 이태영 (2011). 연구데이터와 관련된 OpenURL의 학술서비스 유형 메타태그의 확장에 대한 연구. 정보관리연구, 42(4), 39-58. http://dx.doi.org/10.1633/JIM.2011.42.4.039

김선태 (2012). 관측분야 과학데이터기록의 검색과 발견을 위한 메타데이터 표준요소 선정에 관한 연구. 박사학위논문, 전북대학교 대학원, 문헌정보학과.

안부영, 조희형, 안성수, 박형선 (2005). 생물다양성 데이터교환을 위한 메타데이터 스키마 설계. 한국정보과학회 2005 가을 학술발표논문집, 32(2), 91-93.

이상태 (2006). 과학기술 데이터세트 메타데이터 표준. STISC 과학기술정보표준화위원회.

이혜영, 곽승진 (2007). 지구 관측자료 공유를 위한 메타데이터 연구. 한국문헌정보학회지, 41(2), 257-276. http://dx.doi.org/10.4275/KSLIS.2007.41.2.257

한국과학기술정보연구원 (2011). 분야별 과학데이터 구축 및 활용 현황에 관한 연구. 대전: 한국과학기술정보연구원.

한종엽, 최영준 (2004). 해양전문정보센터의 멀티미디어 메타데이터베이스 및 디지털도서관 통합정보시스템 구현에 관한 연구. 정보관리학회지, 21(4), 5-26. http://dx.doi.org/10.3743/KOSIM.2004.21.4.005

Abrams, S., Rizk-Jackson, A., Kochi, J., & Wittman, N. (2013). Sharing Data-Rich Research Through Repository Layering. OR 2013, PEI, Canada. Retrieved from http://or2013.net/sites/or2013.net/files/slides/OR-2013-Abrams-sharing-data-richresearch.ppt

ACRL Research Planning and Review Committee (2012). 2012 top ten trends in academic libraries. College & Research Libraries News, 311-320. Retrieved from http://crln.acrl.org/content/73/6/311.full.pdf+html

Cox, M. A., & Pinfield, S. (2013). Research data management and libraries: Current activities and future priorities. Journal of Librarianship and Information Science, to be published. http://dx.doi.org/10.1177/0961000613492542

Dietrich, D., Adamus, T., Miner, A., & Steinhart, G. (2012). De-Mystifying the Data Management

Requirements of Research Funders. Issues in Science and Technology Librarianship, 70, summer. doi:10.5062/F44M92G2

Green, T. (2009). We Need Publishing Standards for Datasets and Data Tables. OECD Publishing

Jones, S. (2012). Curation policies and support services of the main UK research funders. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/RC%20policy%20overview%20v2.2.pdf

Juan-le, W., Song-cai, Y., & Chuan-jie, X. (2005). Analysis and Design of Metadata Standard Structure for Geosciences Data Sharing. Institute of Geographical Sciences and Natural Resources, CAS, Beijing 100101, China. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTOTAL-DLGT200501005.htm

Kuipers, T., & Hoeven, J. V. D. (2009). Insight into digital preservation of research output in Europe.

Mabe, M. (2009). An overview of scientific and scholarly journal publishing. The STM report.

McGovern, N. (2013). Trust in Repositories: Building and Measuring Trustworthiness Using TRAC. Open Repositories 2013, Canada. Retrieved from http://or2013.net/sites/or2013.net/files/TrustInRepositories-McGovern_Mar2013_final.doc

Michener, W. K., Allardb, S., Buddenc, A., Cookd, R. B., Douglassb, K., Framee, M., …, & Vieglaisg, D. A. (2012). Participatory design of dataone—enabling cyberinfrastructure for the biological and environmental sciences. Ecological Infomatics, 11(September), 5-15. Retrieved from http://www.sciencedirect.com/science/article/pii/S1574954111000768

Phillips, S., & Koenig, J. (2008). DSpace Administration and Use. Texas Digital Library. Retrieved from http://www.tdl.org/wp-content/uploads/2009/04/DSpaceAdministrationAndUse.pdf

PRC (2010). Access vs. importance. Retrieved from http://www.publishingresearch.net/documents/PRCAccessvsImportanceGlobalNov2010_000.pdf

Science (2011). Challenges and Opportunities. Science, 331(6018), 692-693.

Starr, J., Ashton, J., Barton, A., Elliott, J., Jacquemot-Perbal, M., Karjalainen, M., …, & Ziedorn, F. (2013). DataCite Metadata Schema for the Publication and Citation of Research Data. Retrieved from http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.0.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, U. A., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. doi: 10.1371/journal.pone.0021101

Vempati, D. U., Chung, C., Mader, C., Koleti, A., Datar, N., Vidović, D., Wrobel, D., Erickson, S., Muhlich, L. J., Berriz, G., Benes, H. C., Subramanian, A., Pillal, A., Shamu, E. C., & Schürer, C. S. (2014). Metadata Standard and Data Exchange Specifications to Describe, Model, and Integrate Complex and Diverse High-Throughput Screening Data from the

Library of Integrated Network-based Cellular Signatures (LINCS). Jorunal of Biomolecular Screening, February. doi: 10.1177/1087057114522514

위키백과 (2014, April). 힉스 보손. Retrieved from http://ko.wikipedia.org/wiki/%ED%9E%89%EC%8A%A4_%EB%B3%B4%EC%86%90

Wikipedia (2014a March). E-research. Retrieved from http://en.wikipedia.org/wiki/E-research

Wikipedia (2014b April). e-Science. Retrieved from http://en.wikipedia.org/wiki/E-Science

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

Ahn, B., Cho, H., Ahn, S., & Park, H. (2005). Design of Metadata Schema for Biodiversity Data Exchange. The Korean Institute of Information Scientists and Engineers, Conference.

Han, J., & Choi, Y. (2004). A Study on Planning & Implementation of the Multimedia Meta Database and Digital Library's Integrated Information System for the Oceanographic Information Center. Korea Society for Information Management, 21(4), 5-26. http://dx.doi.org/10.3743/KOSIM.2004.21.4.005

Kim, Suntae (2012). A Study on Standardized Metadata Elements to search and discover observation scientific data. Unpulished doctoral dissertation, University of Chonbuk. Jeonju, Korea.

Kim, Suntae, & Lee, Taeuyoung (2011). A Study on the expansion of Meta-Tag for Research Data in Scholarly Service Type of OpenURL. Journal of Information Management, 42(4), 39-58. http://dx.doi.org/10.1633/JIM.2011.42.4.039

KISTI (2011). Current Status of Constructiong and Utilizing Scientific Data: Survey of 13 Subject Fields.

Lee, H., & Kwak, S. (2007). A Study on Metadata for Sharing the Information of Earth Observation. Korean Society for Library and Information Science, 41(2), 257-276. http://dx.doi.org/10.4275/KSLIS.2007.41.2.257