ISBN 978-294-0762-9

연구데이터 이해와 관리

김선태, 이정훈, 정한민

한 국 과 학 기 술 정 보 연 구 원

ISBN 978-294-0762-9

연구데이터 이해와 관리

김선태, 이정훈, 정한민

한 국 과 학 기 술 정 보 연 구 원

지은이

김선태 / 한국과학기술정보연구원 / 책임연구원
이정훈 / 한국과학기술정보연구원 / 선임연구원

< 목 차 >

<표 차례>

<그림 차례>

1. Data Management Plan 배경

가. 과학 정의와 연구 환경 한계

과학이란 우주에 대한 지식을 모아 테스트 가능한 규칙과 이론으로 압축하는 체계적인 활동이다. 과학의 성공과 신뢰성은 과학자의 의지에 달려 있다. 또한 동료 연구자들에 의한 독립적인 테스트와 복제가 가능하도록 자신의 아이디어와 결과를 공개하는 것에 달려 있다. 이를 위해서 데이터와 연구절차, 필요자원(materials) 에 대한 공개 교환이 요구된다. 이를 통해서 연자는 자신의 과학적 주장을 포기하거나 수정하며 발전 (self-correction) 시킬 수 있다.[1)]

연구자 개인의 지식은 매우 단편적이기 때문에 연구자는 연구과정에서 그리고 연구결과에서 자신이 보고 싶은 것만 바라본다. 이를 'promising findings and nice discoveries'로 표현할 수 있다. 외관상 그럴듯한 이러한 성공적 연구는 재현이 불가능한 경우가 많다. 따라서 과학이 과학답기 위해서는 데이터와 연구절차, 필요자원(materials) 에 대한 공개 교환이 요구되는 것이다. 하지만 연구자 환경은 이러한 요구를 구현하는데 많은 걸림돌을 가지고 있다. 대표적인 특징 네 가지는 다음과 같다.

첫째, 연구자들은 철저한 연구, 엄격한 연구가 창의적이고 혁신적

1) http://www.aps.org/policy/statements/99_6.cfm

인 연구를 방해한다는 잘못된 믿음을 가지고 있다.

둘째, 기대와 다른 설명되지 않은 관찰내용은 철저하게 검증된 후 공표되어야함에도 연구자들은 확인되지 않고 확증할 수 없는 발견을 출판하기에 바쁘다. 따라서, 실험 반복 실패, 합당한 컨트롤(legitimate controls) 사용 실패, 시약 입증(validate reagents), 적정한 통계 테스트 실패가 발생한다.

셋째, 전체 데이터셋 참조 보다는 최상의 실험을 선택하게 되며, 연구 복제가 불가능해 진다. 결국 논문의 주요 결론이 입증되지 못하는 사태가 발생하게 된다.

넷째, 복제 불가능한 연구 출판물이 수백번 인용되고 임상 연구가 진행되는 사례도 있으며, 인용자가 재현시도를 하지 않거나 피인용 논문의 발견(findings)을 변조하며 많은 문제가 야기되기도 한다.[2)]

나. 연구 패러다임 및 도구의 변화

1) 연구 패러다임 변화

2007년 1월11일, Jim Gray 마지막 공개 연설 내용을 토대로, 2009년, The Fourth Paradigm 도서를 Microsoft Research[3)]에서 공개하였다. 마이크로 소프트사 (Microsoft, MS)의 빌 게이츠는

---

2) C. Glenn Begley, John P.A. Ioannidis. 2015. Reproducibility in Science / Improving the Standard for Basic and Preclinical Research. 2015 Circulation Research / https://doi.org/10.1161/CIRCRESAHA.114.303819Circulation Research. 2015;116:116-126 Originally published December 31, 2014

3) MS 부서로서, 1991년 컴퓨팅 기술 개발과 범 지구적 문제를 대학, 정부, 산업계 연구자들의 혁신적 협력을 통해 극복하고자 만들어짐. 1천명 이상의 직원들이 근무(컴퓨터 과학자, 물리학자, 엔지니어, 수학자 등, 컴퓨터분야, 수학분야 노벨상 수상자 영입(Turing Award 수상자, Fields Medal 수상자))

Jim Gray 의 생각에 대해서 “데이터와 소프트웨어가 ‘과학을 한다는 것’의 의미를 재정의하는 것에 대해 새로운 방법으로 사고하도록 함” 이라 말했다.

그림 1 연구 환경 변화

Jim Gray는 그리드컴퓨팅의 대부로서 데이터 집중과학 시대를 미리 예견하고 대용량 데이터의 효율적 조정, 관리, 가시화를 위한 새로운 기술이 필요함을 역설하였다. 책에서는 과학적 발견을 위한 연구의 중심도구가 관찰 > 이론 > 계산 > 데이터로 변하였다고 주장한다. 이를 제 4세대 연구패러다임으로 칭했다.

2) 데이터에 대한 인식 변화

데이터가 국가의 자산이고 기관 및 기업의 자산이라는 인식이 확산되고 있다. 미국의 경우, 과학재단(NSF)이 지원하는 R&D 사업에서 발생한 데이터에 대한 관리와 공동 활용을 위한 지침을 제정하였으며, 호주는 연구수행책임법에서 과학데이터의 관리와 공동 활용에 대한 사항을 규정하고 있다. 영국은 과학데이터의 공동 활용을 위해 데

이터의 관리와 공유를 위한 정책을 수립하였다.

3) 연구 환경의 변화

비록 오픈 사이언스 용어의 역사는 오래되지 않았지만 그 실천은 과학지식과 연구 자원으로의 자유로운 접근을 위한 사회적인 요구에 의해서 출현한 17세기 학술저널(academic journal)부터 시작되었다. 2015년 10월, 현재 시점에서는 연구결과에 대한 자유로운 접근과 배포를 주장하는 그룹과 이에 대해 반대하는 그룹의 주장이 공존하고 있다. 따라서 오픈 데이터, 오픈 액세스와 같은 오픈 사이언스 운동이 주목을 받고 있다.

과학은 데이터의 수집과 분석, 출판, 재분석, 비판(critiquing), 재사용으로 이해될 수 있다. 이러한 과학에 '오픈'이라는 키워드가 붙은 것은 과학의 활성화를 목적으로 한다. 과학 활성화를 방해하는 요인으로는 영리를 추구하는 출판사의 라이선스 정책, 데이터의 사용 제한, 형식화 되어 있지 않은 데이터의 품질, 상업 소프트웨어 사용, 데이터 출판 후 데이터의 오용에 대한 두려움을 들 수 있다. 이러한 요인들을 극복하고자 하는 운동들이 오픈 사이언스 6대 원칙에 포함되어 있으며, 이러한 원칙은 더욱 세분화된 운동으로 확산될 것이다. 예를 들어 상업 소프트웨어 사용과 관련된 '오픈 소스 운동 (open source movement)'의 경우, 데이터의 재생산성을 높이기 위한 구체적인 운동이 전개되고 있다. Nick Barnes가 제시하는 '과학 코드 선언 (Science Code Manifesto)'이 그것이다. Barnes는 과학을 위

한 소프트웨어의 중요성을 역설하며 5대 원칙 (Code, Copyright, Citation, Credit, Curation)을 제시하고 있다.[4)]

4) Science Code Manifesto http://sciencecodemanifesto.org/

2. 데이터

데이터란 facts (사실, 실상, 실제) 의 집합, 재해석 가능한 정보의 표현으로써 다양한 이름으로 표현 된다[5]. 예를 들어, 해양과학 분야에서는 데이터 대신 '자료' 라는 표현을 사용한다. 위키에 따르면, 1946년에 데이터는 'transmittable and storable computer information'이라는 표현으로 처음 사용되었으며, 1954년 'data processing'이라는 용어가 사용되었다. 데이터의 종류로는 microdata, summary data, raw data, primary data, secondary data 등과 같이 구분의 기준에 따라 다양하게 불린다.

가. 데이터 구분

데이터의 상태적 측면을 기준으로 원시 데이터 (raw data), 처리 데이터 (processed data)로 구분되며, 직접생산 여부와 연구목적에 따라 주요 데이터 (primary data), 2차 데이터 (secondary data)로 구분된다. 주요 데이터는 해당 연구를 통해, 직접적인 관찰 및 수집을 통해 생산되는 데이터를 의미한다. 2차 데이터는 이미 출판한 데이터, 과거에 수집된 데이터, 타인이 관찰, 수집한 Primary data, 다른 목적의 데이터[6]를 통칭한다. 데이터 생산 방법에 따르면 관측이나 관찰을 통해 생산되는 observational data, 실험을 통해 생산되는 experimental data, 시뮬레이션을 통해 생산되는 simulation data,

5) http://searchdatamanagement.techtarget.com/definition/data
6) http://www.businessdictionary.com/definition/secondary-data.html

컴파일이나 추출을 통해 생산되는 derived or compiled data, 참조에 사용되는 reference of canonical data 등이 있다.

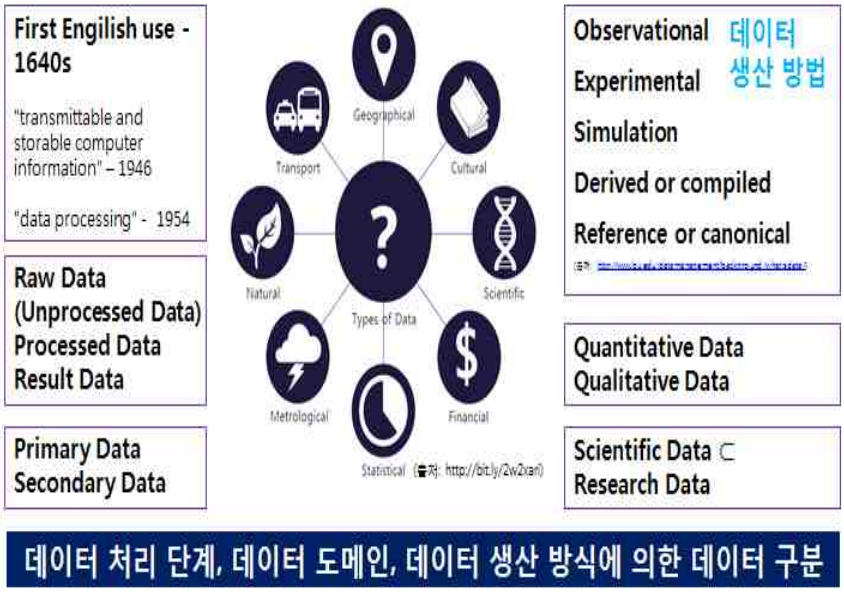


그림 2 데이터 구분

나. 데이터 특징

관측데이터는 현재시점에 생산된 데이터로서 재생산 및 대체 불가능하다. 센서 데이터, 센서 기반 인간 관찰, 설문 결과, 신경 이미지 데이터, 샘플 데이터 등이 있다. 실험데이터는 통제된 조건에서 데이터 가 생산된다. 현재 시점에 실험실에서 주로 생산된다. 재생산이 가능하나 재생산에 소요되는 비용이 높을 수 있다. 유전자 시퀀스, 크로마토그램, 분광데이터, 현미경검사데이터, 환면체 마그네틱 데이

터 등이 있다. 추출 및 컴파일 데이터는 재생산 가능하나 재생산 비용이 고비용일 수 있다. 텍스트 데이터 마이닝 데이터, 추출된 변수 데이터, 컴파일된 데이터베이스, 3D 모델 데이터 등이 있다. 시뮬레이션 데이터는 실제 혹은 이론적 시스템의 행태와 성능을 연구하기 위해 모델로 부터 생산된 결과데이터이다. 모델과 메타데이터가 출력 데이터 보다 중요한 특성을 가지고 있다. 기후 모델, 경제 모델, 생물지구화학 모델 데이터 등이 있다. 레퍼런스 데이터는 검증된 통계 혹은 신체 컬렉션 데이터세트 등을 말한다. 유전체 시퀀스 데이터뱅크, 화학 구조, 통계 데이터, 공간 데이터 포털 등이 있다. 데이터의 양적 측면과 질적 측면을 강조하기 위해서 quantitative data, qualitative data로 구분하기도 한다. 양적 연구 방법(실증적 연구 방법)에 의해 생산되는 양적 데이터는 경험적 자료를 수집하고 계량화하여 사회·문화 현상을 통계적으로 분석하는 연구과정에서 주로 생산된다. 질적 연구 방법(해석적 연구 방법)에 의해 생산되는 질적 데이터는 연구자의 직관적인 통찰로 사회·문화 현상의 의미를 해석하고 이해하려는 연구 방법과정에서 주로 생산된다.

표 3 데이터 구분

구분 기준	데이터 이름	데이터 이름
데이터 처리 단계	Raw Data (Unprocessed Data)	원시 데이터
	Processed Data	처리 데이터
	Result Data	결과 데이터
직접생산 및 연구목적	Primary Data	주요 데이터

구분 기준	데이터 이름	데이터 이름
	Secondary Data	2차 데이터
데이터 생산 방법	Observational	관측·관찰 데이터
	Experimental	실험 데이터
	Simulation	시뮬레이션 데이터
	Derived or compiled	추출·컴파일 데이터
	Reference or canonical	참조 데이터
데이터의 질과 양	Quantitative Data	양적 데이터
	Qualitative Data	질적 데이터
생산 및 활용 분야	Geographical Data	지리 데이터
	Cultural Data	사회문화 데이터
	Scientific Data	과학 데이터
	Financial Data	금융 데이터
	Statistical Data	통계 데이터
	Metrological Data	기상 데이터
	Natural Data	생태 데이터
	Transport Data	교통 데이터
데이터 형태	Record type Data	레코드 기반 데이터
	Graph type Data	그래프 기반 데이터
	Sequential type Data	순차 기반 데이터
데이터 역할	Data	데이터
	Meta Data	데이터에 대한 데이터

한편, 연구과정에서 생산되는 데이터를 연구 데이터 (research data)라 부르며, 이는 과학 데이터 (scientific data)를 포함한다. 한편, 데이터 관련 용어로 데이터 아카이브 (Data Archive), 데이터 보존과 접근 (data preservation and access), 데이터 파일의 구조, 목차, 형식정보를 담고 있는 코드북 (Codebook), 타임 시리즈 (Time Series)와 같이 일정한 시간간격을 둔 순차데이터 용어도 있다.

3. 데이터 세트

데이터 세트의 정의는 분야마다 다양하다. 일반적으로 데이터들의 집합을 의미하며, 메타데이터가 포함되어 있는 경우도 있는데 이는 데이터 설명정보도 데이터세트 정의에 포함될 수 있음을 의미한다. '컴퓨터 처리를 위한 데이터 레코드들의 집합'[7]으로 정의되기도 하며, '웹에서 접근하고 다운로드 할 수 있는 다양한 형태의 데이터 집합[8]'으로 정의되기도 한다. 데이터와 관련된 것들을 담고 있는 주머니를 데이터 세트라 부른다.

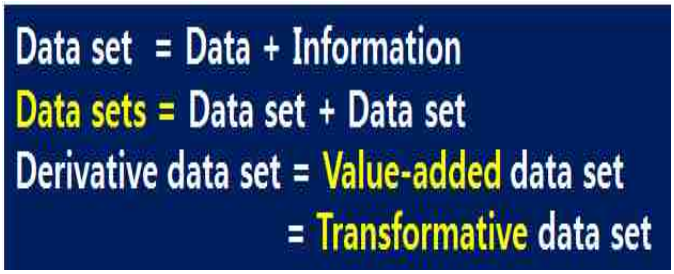


그림 3 데이터세트 정의

원시 데이터, 중간처리 데이터, 그리드, 이미지, 지도, 테이블 등을 포함할 수 있으며, 여러 소스에서 수집되거나 하나의 장치에서 생산될 수 있다. 예를 들어, 다중빔 수중측량기 데이터 집합은 수백개의 swath 데이터 파일을 포함하고 있다[9]. 데이터 세트를 폭넓게 정의하면 데이터와 데이터 관련 정보의 집합을 포함한다. 추출된 데이터 세

7) Dictionary.com
8) W3C Data Catalog Vocabulary
9) http://www.marine-geo.org/help/data_FAQ.php

트는 가치가 부여된 데이터 세트 혹은 변형된 데이터 세트와 같은 개념이다[10].

가. 데이터 세트의 종류

데이터 세트는 구조적 형태의 특징에 따라 레코드 데이터, 그래프 데이터, 순차데이터로 구분될 수 있다. 데이터 행렬의 경우는 행렬의 값들이 모두 숫자로 구성된 경우가 대부분이다. 다차원 공간의 포인트로 간주할 수 있다. 텀 벡터는 문서들의 텀들과 문서에 나타난 빈도수 정보의 리스트 형태이다. 각각의 문서는 하나의 텀 벡터가 된다. 트랜잭션 데이터는 하나의 레코드(트랜젝션)가 여러 아이템을 포함하는 특징을 갖는다. 고객이 구매한 물품 리스트 데이터 등이 있다.

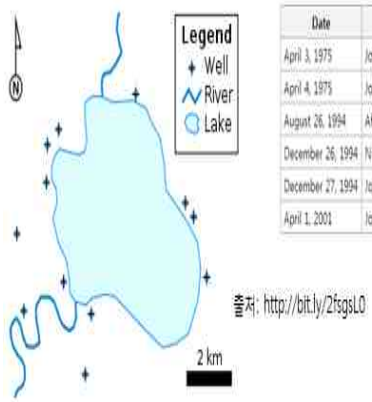
10) Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, 1999, p. 15

그림 4 레코드유형 데이터세트

그래프 기반 데이터는 화합물의 입체적 구조를 표현한 형태를 갖는다. RDF (Resource Description Framework) 형태의 데이터도 이에 해당한다. 순차 기반 데이터는 주로 Geospatial Data (Spatial Data와 attribute data), Temporal Data, Sequential Data, Genetic Sequence Data 등으로 구분된다.

공간 데이터는 위치에 해당하는 정보로서 도로의 모양이나 좌표에 대한 정보를 의미한다. 여기에 속성 데이터 (attribute data) 로서 도로의 속성에 대한 정보 (이름, 길이, 속도 제한, 혹은 방향등의 정보)가 포함된 것이 지리 공간 데이터이다. Temporal 데이터는 당시의 상태를 표현한 데이터로서 많은 소스 (수동 입력, 관측센서, 시뮬레이션 모델 등)로 부터 획득된다. 예를 들어, 1990년 홍콩의 토시사용 패턴, 2009년 7월1일 호놀룰루 총 강수량, 해양 포유류 위치 가시화,

도시 인구 증가의 이해, 특정 질병으로 인한 사망자수 연구, 해양 기후 및 날씨 패턴 변화 데이터 등이 있다.



Date	Real world event	Database Action	What the database shows
April 3, 1975	John is born	Nothing	There is no person called John Doe
April 4, 1975	John's father officially reports John's birth	Inserted Person(John Doe, Smallville)	John Doe lives in Smallville
August 26, 1994	After graduation, John moves to Bigtown, but forgets to register his new address	Nothing	John Doe lives in Smallville
December 26, 1994	Nothing	Nothing	John Doe lives in Smallville
December 27, 1994	John registers his new address	Updated Person(John Doe, Bigtown)	John Doe lives in Bigtown
April 1, 2001	John dies	Deleted Person(John Doe)	There is no person called John Doe

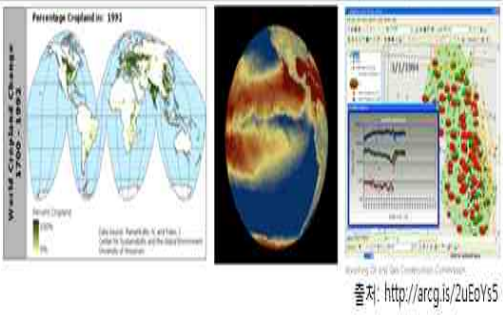


그림 5 지리 공간유형 데이터세트

Sequential Data는 복수의 아이템 세트로 구성되어 있으며, 각각의 아이템 세트는 여러 개의 아이템들을 가지고 있다. 같은 아이템 세트에 존재하는 아이템들은 동일한 타임스탬프를 갖는 특징이 있다.

나. 데이터 품질

데이터 품질은 매우 중요하다. 데이터를 가공하는 과정은 고품질 데이터를 생산하기 위한 과정과 같다. 데이터 가공은 데이터 정제 (data cleaning) 로도 불린다. 데이터를 정제하는 과정에서는 데이터 품질에 영향을 미치는 노이즈 (잡음), 이상치 (outliers), 중복 데이터

(duplicate data), 누락 값 (missing values) 등을 적합한 방법으로 처리해야 한다.

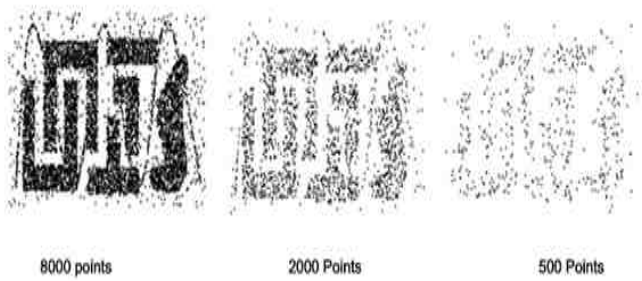


그림 6 샘플링 예시 http://bit.ly/2unU33Z

노이즈는 원본 값의 변경을 의미한다. 음성의 왜곡이나 TV스크린의 흔들림과 같은 것을 말한다. 이상치는 데이터 집합 내에 다른 값들과 상당히 다른 특징을 갖는 값을 의미한다. 중복 데이터는 동일한 값이 데이터 집합 내에서 2회 이상 출현하는 것을 의미한다. 누락 값은 데이터 값이 채워지지 않은 공백 값을 의미한다. 각각의 품질 문제를 해결은 품질문제 발생원인과 데이터 정제 목적에 따라 다르다. 예를 들어, 국민 연간소득 데이터의 경우, 누락 값은 다양한 이유로 발생할 수 있다. 데이터 제공을 동의하지 않았을 경우 발생할 수 있다. 또한 아이들의 경우, 연간 소득 데이터가 없기 때문이다. 이러한 누락 값을 처리하는 방법 또한 다양하다. 레코드를 삭제하거나 추측되는 값으로 채울 수 있다. 레코드는 그대로 유지하되 분석에서 제외하는 방법도 있다. 또한 가중치를 판단해서 값을 채우는 방법도 있다. 이상과 같은 데이터 품질 문제를 처리하는 방법으로는 집계

(aggregation), 샘플링 (sampling), 차원 축소 (dimensionality reduction), 특징 선택 및 추출 (feature selection & extraction) 등이 있다.

집계는 데이터 속성 혹은 객체 수를 줄이는 방법이다. 이를 통해서 분석규모를 변경할 수 있다. 보다 정데된 안정적인 데이터를 확보하는 방법 중 하나이다. 집계 방법으로는 여러 개의 속성을 하나의 속성으로 통합하거나 여러 개의 객체를 하나의 객체로 통합하는 방법이 있다. 샘플링은 관심 있는 모든 데이터 확보 및 분석에 높은 비용이 발생하고 많은 시간이 소요되기 때문에 데이터 마이닝 과정에서 사용되기도 한다.

차원의 크기는 특징(feature)의 개수를 의미한다. 차원축소는 데이터의 의미를 제대로 표현하는 특징을 추려내는 것이다. 차원이 증가하면 그것을 표현하기 위한 데이터 크기가 기하급수적으로 필요하기 때문에 고차원 데이터들은 의미를 제대로 표현하기 어렵다.

데이터의 차원을 줄이는 방법으로 특징 선택과 특징 추출이 있다. 특징 선택은 모든 특징의 부분 집합을 선택해서 간결한 특징 집합을 만드는 것이다. 즉, 원본 데이터에서 불필요한 특징들(변수들)을 제거하는 것이다. 예를 들어, 변수 X와 변수 Y의 특징이 점프 높이 결과 예측에 영향이 없다고 생각한다면 전체 특징 집합에서 해당 특징들을 제거해 간결한 특징 집합을 만드는 것이다. 특징 추출은 원본 특징들의 조합으로 새로운 특징을 생성하는 것이다. 예를 들어, 주성분분석(Principal Component Analysis)은 데이터로부터 직교 주축을 찾고

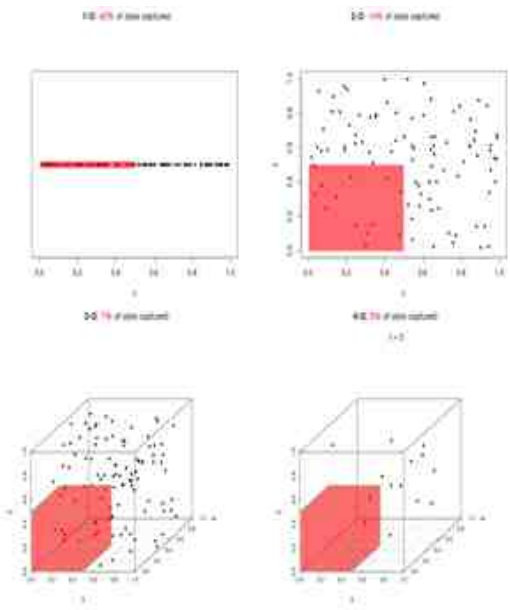


그림 7 차원변화에 따른 필요데이터양 (http://bit.ly/2uLGeLT)

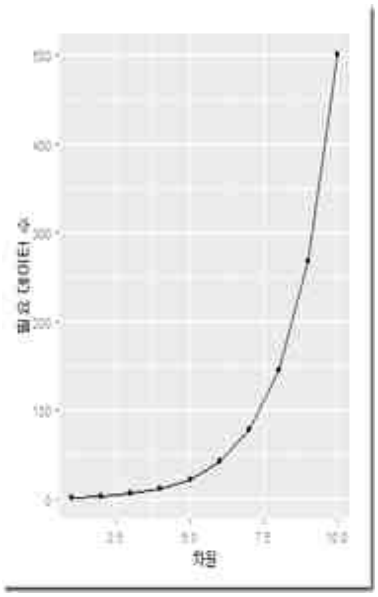


그림 8 차원변화에 따른 필요 데이터양 (http://bit.ly/2vTYabG)

모든 데이터를 해당 축에 투영시킨다. 이 경우, 원본 데이터를 투영된 데이터로 만드는 투영 함수는 결국 원본 특징들의 선형 결합으로 이루어진 새로운 특징을 만드는 것이다.

4. 메타데이터

데이터는 오프라인과 온라인상에서 접근 가능한 모든 정보자원(information resources, item, object)을 의미한다. 정보자원은 도서와 같은 물리적인 자원뿐만 아니라 컴퓨터 파일, 문서, 이미지, 데이터세트와 같은 디지털 자원을 모두 의미한다. 메타데이터는 '정보자원에 대한 데이터'이다. '데이터에 대한 데이터'가 메타데이터의 보편적 정으로 사용되는 이유이다. 이상을 종합해 볼 때 메타데이터란 정보자원이나 정보자원의 묶음(collection, object)을 설명(기술, describe)한 데이터를 말한다. 연구 기록 (research records)은 종이나 전자파일 형식으로 존재하는 것으로서 데이터와 정보(자료)를 포함한 문서다. 연구 과정과 관련된 기록으로서 (전자)메일, 프로젝트 파일, 연구비 신청서, 윤리신청서, 저작권 협약서, 기술보고서, 연구보고서, 실험노트북, 연구저널, 마스터리스트, 동의서, 연구자 참여정보 등을 포함한다. 이러한 연구 기록도 관리와 활용을 위해서 메타데이터가 필요하다.

메타데이터의 대표적인 예로 도서관의 목록(catalog)을 들 수 있다. 도서관 목록은 도서나 저널, 그 밖에 도서관이 보유(소장)하고 있는 정보자원에 대한 정보로 구성되어 있다. 예를 들어, 어떤 도서(책)의 제목이나 저자, 출판사, 출판년도 등의 정보가 해당 도서(정보자원)에 대한 정보로서 목록의 내용을 구성하고 있다. 이렇듯 책이라는 정보자원을 설명하는 제목, 저자, 출판사 등을 메타데이터 요소

(metadata elements, property, 속성)라 하며, 그 값을 메타데이터 요소의 속성값(속성치)이라 한다.

그림 9 디지털 카메라의 메타데이터 설정
사진출처: http://bit.ly/1JD2SJJ

사진이라는 정보자원(아이템, 데이터라고도 함)을 하나의 예로 더 살펴보면, 사진을 찍은 사람, 사진 속 인물은 누구인지, 사진을 언제 찍었는지, 어디에서 찍었는지, 어떤 카메라로 찍었는지 등의 정보는 사진이라는 정보자원에 대한 데이터(메타데이터)이다. 이 가이드의 '메타데이터의 종류' 부분에서 자세히 언급되겠지만, 메타데이터는 사람이 직접 입력해야 하는 것도 있고, 정보자원을 생산하는 생산장비나 정보자원을 관리하는 시스템 등에서 자동적으로 생산되는 것도 있다. 사진의 경우, '사진 속 인물이 누구인지'는 사람이 직접 만들어야 하는 메타데이터 이지만, '사진을 찍은 사람'과 '사진을 언제 찍었는지' 등의 정보는 카메라에서 자동으로 생산해 낼 수 있는 메타데이터이다. 사진의 해상도나 사진이 찍힌 위치정보(좌표값) 등도 후자에 해

당된다.

정보자원의 묶음(collection)이란 여러 개의 정보자원을 하나로 모은 것을 말한다. 예를 들어, 디지털 사진(정보자원, object, item, resource)을 여러 장을 CD(compact disc)에 저장했을 때, 이 한 장의 CD가 컬렉션이 될 수 있으며, 인화된 사진을 모아놓은 앨범도 컬렉션이 될 수 있다. 물론 디지털 사진을 모아 놓은 하나의 컴퓨터 폴더(folder)도 컬렉션이 될 수 있다. 폴더가 폴더를 포함할 수 있는 것과 같이 컬렉션은 컬렉션을 포함할 수 있다.

이러한 컬렉션(정보자원의 묶음)을 설명하기 위해서도 메타데이터가 필요하다. 예를 들어, 컬렉션 이름, 컬렉션의 크기(size), 누가 컬렉션을 만들었는지? 언제 만들었는지? 등의 정보(메타데이터)는 컬렉션 자체가 아니라 컬렉션을 설명(기술)하는 데이터이다. 이러한 데이터가 컬렉션을 설명하는 메타데이터이다.

하나의 사진파일도 '사진 이름(title)'이 있을 수 있으며, 폴더나 CD도 '폴더 이름(title)'이나 'CD 이름(title)'이 있을 수 있다. 'title'이라는 동일한 메타데이터 속성으로 각각의 속성치(속성값)을 기술(describe)할 수 있다.

가. 메타데이터의 필요성

정보자원에 따라서 메타데이터의 중요성에는 다소 차이가 있다. 하지만 메타데이터의 공통적 기능은 정보자원의 유형과 관계없이 거의 동일하다. 이는 메타데이터의 고유기능이 존재함을 의미한다. 메타데이터가 사용되는 용도에 따라 대표적인 메타데이터 스키마가 존재하는 것은 메타데이터의 고유기능이 세분화 될 수 있음을 의미한다.

메타데이터는 정보자원을 설명(기술)한 데이터이다. 따라서 메타데이터를 통해서 데이터가 존재하게 된 관련정보(context information, 상황정보)를 알 수 있다. 따라서 사진이나, 문서, 동영상, 도서와 같은 정보자원과 달리, 단순한 수치 데이터만을 담고 있는 정보자원의 경우 메타데이터의 효용성이 더욱 크다. 예를 들어, 특정 실험 장치로 데이터(정보자원, 원시데이터)를 획득했을 때, 실험의 조건이나 장치이름, 장치버전, 장치에 탑재된 소프트웨어 이름과 버전 등의 정보는 데이터가 생산되는 시점의 상황 정보를 알 수 있다. 따라서 데이터를 이해하는데 정보를 제공할 수 있으며, 데이터의 재사용을 보장할 수 있다.

나. 메타데이터 요소의 유형

메타데이터 요소들의 쓰임새에 따른 유형별 구분은 다음과 같다. 사진(정보자원)을 가지고 사진과 관련된 메타데이터 요소를 살펴보면 다음과 같다.

- Descriptive Metadata
  - (사진) 데이터를 설명하기 위한 메타데이터
  - 메타데이터 예: 사진을 찍은 사람(photographer), 사진의 주제(subject), 사진을 찍은 일시(date, time)

- Technical Metadata
  - (사진) 데이터의 기술적 특징을 담기 위한 메타데이터
  - 메타데이터 예: 카메라 유형(type), 사진 크기(dimensions), 해상도(resolution)

- Access or rights metadata
  - (사진) 데이터의 접근과 이용(라이선스) 정보를 담기 위한 메타데이터
  - 메타데이터 예: 접근 범위(access), 접근 조건(right, license)

- Preservation metadata
  - (사진) 데이터의 지속적 접근과 활용을 보장하기 위한 메타데이터
  - 메타데이터 예: 데이터 변환일(date), 데이터 버전(version), 운영시스템 종류(os)

다. 메타데이터 스키마

상기의 메타데이터 요소들(elements, terms)을 조합하여 특정한 목적의 메타데이터 요소세트를 구성한 것이 메타데이터 스키마이다. 메타데이터 스키마는 정보자원을 위한 메타데이터 요소들의 구성 순서와 반복사용 여부 및 횟수, 메타데이터 요소 사용에 있어 필수 및 권고, 재량 등의 설명, 메타데이터 요소 값의 형식 등의 정보를 갖는다.

메타데이터 스키마는 그 목적에 따라 다양한 표준이 존재한다.

- 정보자원의 검색, 관리, 서비스하기 위한 스키마
  - 정보자원을 생산하는 주제 분야별로 다양한 스키마가 존재한다. 지구과학 분야에서 사용하는 DIF와 사회과학분야에서 사용하는 DDI, 도서관 분야에서 사용하는 MARC-XML 등이 대표적인 스키마이다.

- 메타데이터(정보자원)을 교환하기 위한 스키마
  - METS는 표준화된 형식으로 메타데이터를 교환하기 위한 대표적인 스키마이다.

- 정보자원을 출판하기 위한 스키마
  - DataCite나 CrossRef에서 연구 데이터(research data)와 논문의 출판을 위한 메타데이터 스키마를 제안하고 있다.

- 정보자원을 관리하는 리포지터리(repository) 스키마 등
  - DCMI에서는 웹 환경에서 접근 가능한 모든 것을 자원(resource)으로 정의하고 있다. 따라서 정보자원을 수집, 관리, 배포하는 기관 리포지터리(institutional repository, IR) 또한 자원이다. 따라서 IR에 대한 설명(IR 이름, IR 운영기관 등)도 메타데이터로 기술될 수 있다. IR을 설명하기 위한 메타데이터 스키마(메타데이터 요소를 어떻게 구성할지 정의한 것) 또한 다양하다. 호주의 RIF-CS나 re3data.org가 IR을 설명하기 위한 대표적인 메타데이터 스키마이다.

5. 연구 데이터 : Research Data

미국 행정 관리 예산국 (OMB)에 따르면, 연구 데이터란 과학 커뮤니티에서 연구발견을 검증하는데 필요하다고 인정되는 '기록된 사실 자료 (recorded factual material)'로 정의 할 수 있다. 연구의 재현과 검증을 위해 필요한 것이 연구 데이터라는 이야기다. 연구 데이터 정의는 대동소이하다, 일부 정보자원 유형의 포함 여부가 상이할 뿐이다. 아래 내용은 연구 데이터 집합에 포함되는 정보자원의 요소를 수학기호로 표현한 것이다. 집합의 의미는 '{'과 '}'으로, 같은 의미는 '='으로 표기하였다. 포함 관계는 '⊂', 원소는 '∃', 원소가 아님은 '∄'으로 표기하였다.

다음은 Datacite.org[11]에서 바라보는 연구 데이터 정의이다.

Research Content = Research Objects

Research Data ⊂ Research Objects

Research Objects ∃ Workflows

Research Data ∄ Workflows ==> 이 부분은 보스턴 대학의 정의와 상이

Research Objects ∃ Standards

Research Data ∄ Standards

Research Data ∃ Dataset

11) 국제적인 컨소시엄으로서 2016년 11월 현재, 20개국 이상에서 여러 데이터 센터, 도서관, 정부 조직, 대학들이 컨소시엄 멤버로 참여중이다

다음은 퀸스랜드 대학의 연구 데이터 정의이다.

{ facts, observations, images } ⊂ Research Data
{ computer program results, recordings } ⊂ Research Data
{ measurements, experiences } ⊂ Research Data

다음은 멜본 대학의 연구 데이터 정의이다. 멜본 대학은 연구 데이터 뿐 아니라, 연구 컬렉션과 연구 레코드에 대한 정의도 제공한다.

{ facts, observations or experiences, laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses } ⊂ Research data
{ slides; artefacts; specimens; samples } ⊂ Research collections
{ electronic mail as well as paper-based correspondence; project files; grant applications; ethics applications; authorship agreements; technical reports; research reports; laboratory notebooks or research journals; master lists; signed consent forms; and information sheets for research participants } ⊂ Research records

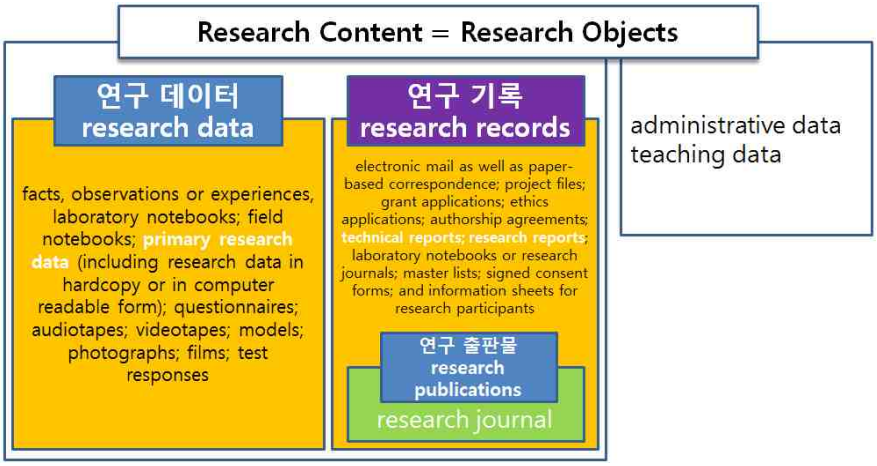


그림 10 연구 데이터와 연구 기록, 연구 출판물 구분

이렇듯 연구 데이터에 대한 정의는 정의하는 주체에 따라 다양하지만 거의 대동소이하다. 보스턴 대학의 경우는 연구 데이터 정의 및 자격요건을 보다 명확하게 제시한다. 연구 데이터를 '사실적 기록으로써 연구 발견을 재현 (Reproducible science), 검증 (Validating research findings) 하는데 사용되는 데이터'로 정의하고 있다. 한편, 연구 데이터 정의를 반어적으로 제공하는 것이 인상적이다. 연구 데이터가 아닌 데이터를 제시함으로써 연구 데이터를 명확하게 정의하고 있다. 관리 데이터 (administrative data), 교수학습 데이터 (teaching data), 연구 출판물 (research publications)은 연구데이터가 아니라 정의하고 있다. 하지만 Datacite와 ANDS[12]의 경우는 연구 출판물을 연구 데이터 범주에 포함하고 있다. Datacite에서는 Methodologies와 workflows를 연구 데이터 범주에 포함하지 않는

12) 호주의 데이터 거버넌스 서비스 http://www.ands.org.au/

반면, 보스턴 대학은 포함한다. 한편, 그리피스 대학은 설문조사, 녹음 자료를 연구 데이터로 분류하지 않고 'primary materials' 로 분류하고 있다는 점이 특이하다. 필자는 보스턴 대학의 반어적 구분에 일부 동의하지만 연구 데이터를 다음과 같이 정의할 수 있다.

연구 데이터란 연구 과정에서 수집, 생산된 모든 정보자원으로써, 연구 결과를 복제, 재현, 증명하는데 필요한 데이터이다. 연구 기록(research records)은 연구과정 및 결과의 기록으로써 연구 출판물(research publications)을 포함한다. ANDS 및 Datacite에서도 연구 출판물이 연구 데이터 범주에 포함되나 연구 출판물의 경우, 정보유통 생태계가 안정적이기 때문에 연구 출판물을 제외한 협의의 연구 데이터를 정의했었다. 따라서 연구 기록을 연구 데이터로 포함할 수 있으나 협의의 연구 데이터 정의에서는 제외시키는 것이 보다 명확하다.

연구데이터와 과학데이터 관계연구 데이터 중 과학 활동을 통해 수집, 생산 된 데이터를 과학 데이터라 부른다. 공공 데이터란 공공재원을 사용하여 수행된 연구 및 조사 활동을 통해 수집, 생산된 데이터이다. 따라서 공공재원으로 수집, 생산된 과학 데이터는 공공 데이터로 분류할 수 있다.

과학데이터 정의는 다양하다. Cheng (2006)은 과학기술 활동의 결

과로서 관측 (Observation), 관찰 (Monitoring), 조사 (Investigation), 실험 (Experiment), 연구 분석 (Research Analysis), 계산 (Computation) 등의 활동을 통해 생성된 데이터[13] 로 정의하고 있다. OECD (2006)는 과학 연구수행을 위한 주요한 원천으로 사용하는 사실적인 기록 (숫자, 문자정보, 이미지 및 소리)로 정의한다. CCSDS (2002)는 전달, 해석 및 가공에 적합하도록 형식을 갖춘, 재해석이 가능한 정보의 표현으로 정의한다. 김선태 (2011)는 연구자의 연구 활동 과정 중 생성되는 다양한 유형 의 사실적 기록. 즉, 연구활동을 통하여 생산 된 연구활동의 기록물로서 관측, 감시, 조사, 실험, 분석, 계산 등의 과정을 통하여 생산된 문자, 이미지, 오디오, 동영상 등의 아날로그 및 디지털 형식을 포괄하는 데이터[14] 로 정의하고 있다.

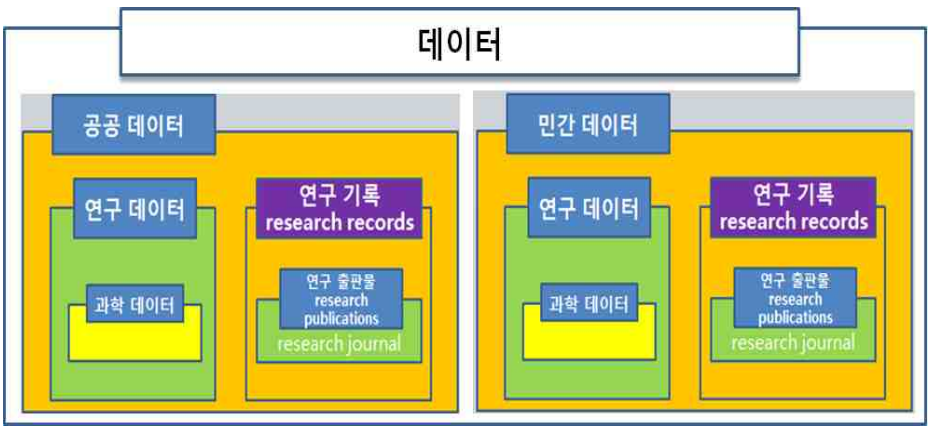


그림 11 공공 데이터와 민간 데이터 구분

13) Cheng, Jinpei. 2006. Strategies for Preservation of and Open Access to Scientific Data in China: Summary of a Workshop

14) 김선태(2011),「과학데이터 보존 및 활용모델에 관한 연구」 http://scholar.ndsl.kr/schArticleDetail.do?cn=JAKO201013351026193

과학데이터 사례로는 실험데이터, 통계데이터, 단백질 구조이미지, 생물의 표본 자료, 천문학의 분광관측(spectral survey) 자료 등을 들 수 있다. 과학데이터 유형은 연구 분야 및 연구방법, 관측장비, 실험장비, 분석방법 등에 따라 다양하다. 소량의 통계데이터 부터 가속기를 통해 매년 16페타바이트씩 생산되는 대용량 미립자 충돌 데이터까지 규모와 형태 적인 측면에서 매우 다양한 특징을 갖는다. 주로 수치정보, 공간정보, 도표정보, 문서 등의 형태를 가지며, 지구관측 및 환경 분야의 데이터는 주로 관측데이터로서 공간 및 수치정보와 이미지 정보이다. 사회과학 분야의 데이터는 주로 설문조사를 통한 통계데이터 형태를 취하며, 컴퓨터과학 분야의 데이터는 주로 도표 또는 수치정보 형태이다.

6. 데이터 관리 계획서

데이터 관리 계획서 (Data Management Plan, DMP)는 연구 과정이나 연구 프로젝트 종료 후에 데이터가 어떻게 취급되는지 기술한 문서이다. 프로젝트가 시작되기 전에 데이터 관리, 메타데이터 생성, 데이터 보존, 데이터 분석과 같은 다양한 측면을 미리 고민하도록 하며, 현재와 미래에 데이터가 잘 관리되도록 하고 데이터 보존을 위한 준비가 가능하도록 한다. 따라서 DMP의 기대효과는 다음과 같다.

- 데이터 수집 전, DMP 준비 : 정확한 형식의 데이터 수집, 제대로 된 데이터 관리, 충실한 데이터 설명이 가능
- 데이터의 재구성, 포맷변환, 그리고 데이터에 대한 상세내역을 기억하기 위한 노력이 불필요
- 데이터 수집자 및 타 연구자들이 잘 설명된 데이터에 대해 이해할 수 있어 재사용 가능

DMP는 다음 그림과 같이 7개 범주로 나누어 기술될 수 있다. DMPonline의 경우도 다음의 7개 범주를 세분화한 13개 항목을 기술하도록 되어있다.

그림 12 데이터 관리 계획서 구성 항목

가. 데이터 관리 계획서 항목별 특징

1) 데이터 컬렉션

데이터 컬렉션 항목에 기술되어야 하는 정보는 다음과 같다. 어떤 유형(type)과 형식(format)의 데이터를 수집할 것인가? 데이터의 장기적 접근 및 재사용 가능한 유형과 형식을 고려하여 작성하도록 권고되어야 한다. 또한 수집이나 생산할 데이터의 크기는 어느 정도 되는지, 연구 프로젝트에서 재사용할 데이터가 있다면 무엇인지 기술되도록 해야 한다. 이 외에도 연구 커뮤니티 데이터 표준 여부, 폴더 구조, 파일 명명법, 버전 컨트롤, 수집 일관성, 품질제어 방법 등이 기술 되어야 한다.

2) 문서화와 메타데이터

문서화와 메타데이터 항목은 ‘미래에 제3자의 데이터 이해 및 재사용’을 염두에 두고 기술해야 한다. 데이터 검색을 위한 기본항목으로써 데이터 생산자, 데이터 공헌자, 데이터 이름, 데이터 생산일, 데이터 접근 조건 등이 반드시 기술되어야 하며, 선택적으로 데이터 생산 방법, 분석 방법, 절차 정보, 변수 정의, 용어집, 측정 단위, 가정(가설), 데이터 형식 및 유형 등이 기술되어야 한다. 물론 가능한 커뮤니티 표준을 사용하여 권고되어야 한다.

3) 윤리와 법률 준수

윤리적 이슈는 데이터 저장 방법, 데이터 접근 대상 및 이용방법, 데이터 보존 기간에 영향을 준다. 따라서 윤리와 법률 준수 부분은 피시험자가 포함된 연구의 경우, 데이터 보존과 공유를 위한 동의 확보 여부를 반드시 기술하도록 해야 하며, 필요한 경우, 개인정보 보호 방법 (익명처리 등)과 개인정보 취급 방법 (안전한 저장 및 전송 방법)을 기술하도록 해야 한다. 윤리적 문제를 다룬다는 것은 데이터의 익명처리, 부서나 조직의 윤리위원회로 이첩, 공식적 동의확보 등을 포함한다. 한편, 수집되거나 생산된 데이터와 관련하여 저작권과 지적재산권을 누가 소유하는가? 데이터 (재)사용 권한은 누가 갖는가? 등에 대해 명확한 내용이 요구된다. 특히, 다자협력 프로젝트의 경우, 지적재산권 소유는 컨소시엄 협약에서 다루어져야 한다. 저작권과 지적재산권에 대한 다양한 이해그룹 (연구비지원기관, 소속 기관, 부서, 그룹 정책) 의 정책이 참고 될 수 있으며, 제 3자 데이터

사용 허가권 및 데이터 공유와 관련된 제약조건을 고려해야 한다.

4) 저장소와 백업

저장소와 백업 항목의 경우, 데이터의 백업 주기, 백업 위치, 복사본 수 등이 기술되어야 한다. 노트북, 컴퓨터 하드디스크, 외장디바이스 저장을 지양하도록 해야 하며, IT팀이 제공하는 강력한 저장장치 사용을 권고해야 한다. 수동에 의한 백업을 지양하고 IT서비스로 제공되는 자동백업을 지향하도록 해야 한다. 제 3서비스로 백업서비스를 사용하는 경우 연구비지원기관, 소속기관, 부서 및 그룹 정책과의 충돌을 검토해야 한다. 특히, 데이터의 보관 위치 혹은 민감 데이터 보호 등과 관련해서 충분한 검토가 필요하다. 데이터 보안 위협 요소 도출 및 처리방안, 데이터 보안을 유지하기 위한 접근 제어 방법, 협력자들의 안전한 데이터 접근 방법, 야외에서 생산 및 수집되는 데이터의 안전한 전송 등이 기술되어야 한다. 민감 정보가 포함된 데이터의 경우, 합리적인 보안 수단을 제시할 필요가 있으며, ISO 27001과 같은 공식적 표준 준수와 같은 표기가 필요하다.

5) 데이터 선정과 보존

데이터 선정과 보존 항목에서는 계약, 법률, 규제와 관련하여 보존 혹은 파기 되어야 하는 데이터, 유지 되어야하는 데이터 결정 방법, 미래에 데이터를 사용할 예측 가능한 연구, 데이터 보유 및 보존 기

간, 잠재적 재사용 가치 등이 기술되어야 한다. 특히, 연구 발견 검증 및 새로운 연구 수행 또는 교육을 위해서 데이터가 어떻게 재사용될 수 있는지 기술되어야 한다. 한편, 데이터 공유와 보존을 위해서 파일형식 변환과 같은 부수적인 노력이 요구될 수 있음을 고려해야 한다. 보존의 경우, 데이터 장기 보존 계획이 선행되어야 한다. 리포지터리 혹은 아카이브 고려 시, 데이터 보존비용을 함께 검토해야 한다. 프로젝트가 종료된 후 장기 보존이 필요한 데이터세트를 어떻게 보존할 것인가? 데이터 공유 및 아카이빙을 위한 준비작업과 문서작업 계획을 어떻게 진행할지에 대한 고민이 필요하다. 안정적인 리포지터리를 사용하지 않을 경우, 프로젝트가 종료되어도 데이터가 효율적으로 관리 되도록 자원들과 시스템이 어떻게 구축되는지를 기술해야 한다.

6) 데이터 공유

데이터 공유 항목에서는 잠재적 이용자의 데이터 검색을 고려하여 데이터 공유 대상과 공유 조건을 검토해야 한다. 또한 데이터 공유채널로서 리포지터리를 통해 데이터를 공유할 것인지, 데이터 사용 요구를 직접 처리할 것인지 또한 검토해야 한다. 데이터 사용을 언제 가능하도록 할 것인지와 데이터 식별자를 사용할지 여부에 대한 고민도 필요하다. 데이터 재사용 시 요구하고자 하는 인용방법을 기술할 필요도 있다. 한편, 데이터 공유와 관련된 제약조건은 무엇인지, 제약조건을 극복하고 최소화하기 위한 조치로 무엇이 필요한지, 데이터

배타적 사용권 기간과 이유, 데이터 공유 협약의 필요여부 등을 고민해야한다. 또한 비공개 협약이 기밀데이터를 충분히 보호해주는지도 검토가 필요하다.

7) 책임과 자원

책임과 자원 항목에서는 DMP 시행/검토/수정 책임자와 데이터 관리 활동 책임자가 기술되어야 한다. 특히, 협력 연구 프로젝트의 경우 데이터 관리 책임 소재가 분명하게 기술되어야 한다. 데이터 소유권 및 연구데이터 관리책임이 연구 파트너 사이의 컨소시엄 협약 혹은 계약에 포함되어있는지 확인이 필요하다. 구체적으로는 데이터 획득, 메타데이터 생산, 데이터 품질, 저장 및 백업, 데이터 아카이빙과 데이터 공유 등과 관련된 역할과 책임 명시가 분명하게 기술되어야 한다. 또한 관련된 정책 준수 여부 확인자를 가능한 한 이름으로 명시한다. 데이터 관리 계획을 이행하기 위해 요구되는 자원으로써 전문가, 하드웨어, 소프트웨어 등을 명시할 수 있으며, 데이터 리포지터리 이용을 위한 비용 또한 기술될 수 있다.

이상을 요약해 정리하면 다음 표와 같다.

표 5 데이터 관리 계획서 세부구성 항목

구분	영문 항목	국문 항목
프로젝트 정보	Project title	프로젝트 이름
	Primary research organisation	연구기관

구분	영문 항목	국문 항목
	Funding organisation	연구비 지원기관
DMP 기본정보	Plan name	데이터 관리 계획서 이름
	ID	관련 아이디
	Grant number	연구비 참조 번호
	Principal Investigator/Researcher	연구책임자 이름
	Principal Investigator/Researcher ID	연구책임자 아이디
	Plan data contact	연락처 정보
	Description	데이터 관리 계획서 개요
데이터 수집 및 생산 (DataCollection)	Data Type	데이터 유형
	Data Format	데이터 형식
	Data Volume	데이터 크기
문서작업 및 메타데이터 (Documentationand Metadata)	Documentation	데이터 문서
	Metadata	메타데이터
윤리 및 법률준수 (Ethics and Legal Compliance)	consent for data preservation and sharing	동의서확보
	protect the identity of participants	개인정보
	sensitive data	민감정보
	copyright	저작권
	Intellectual Property Rights (IPR)	지적재산권
저장 및 백업 (Storage and Backup)	Storage and Backup	저장 및 백업 방법
	Access and Security	접근 및 보안
데이터 선정 및 보존 (Selection and Preservation)	Selection	데이터 선정
	Preservation	데이터 보존

구분	영문 항목	국문 항목
데이터 공유 (Data Sharing)	Data Sharing Method	데이터 공유 방법
	Restrictions on data sharing required	데이터 공유 제약
책임 및 자원 (Responsibilities and Resources)	Responsibilities	데이터 관리 책임
	Resources	DMP 이행을 위한 자원

나. 데이터 관리 계획서 가이드

여기에서는 데이터 관리 계획서 항목별로 기술되도록 권장되어야 할 사항을 살펴본다.

1) 데이터 컬렉션

- 데이터 유형(type)과 형식(format) : 데이터의 장기적 접근과 재사용 가능한 데이터 유형과 형식을 고려하여 작성
- 수집이나 생산할 데이터 크기 (Volume)
- 연구 프로젝트에서 재사용할 데이터 (Secondary Data)
- 연구 커뮤니티 데이터 표준
- 폴더 구조 및 파일 명명법
- 버전 컨트롤 방법
- 수집 일관성 유지 방법

- 품질제어 방법 등

2) 문서화와 메타데이터

- 필수 항목: 데이터 생산자, 데이터 공헌자, 데이터 이름, 데이터 생산일, 데이터 접근 조건
- 선택 항목: 데이터 생산 방법, 분석 방법, 절차 정보, 변수 정의, 용어집, 측정 단위, 가정(가설), 데이터 형식 및 유형 등
- 커뮤니티 표준 사용 권고

3) 윤리와 법률 준수

- (피시험자가 포함된 경우) 데이터 보존과 공유를 위한 동의확보 정보
- (필요한 경우) 개인정보 보호 방법 (익명처리 등)
- 개인정보 취급 방법 (안전한 저장 및 전송 방법)
- 저작권과 지적재산권 소유 정보
- 데이터 (재)사용을 위한 권한 정보
- 제 3자 데이터 사용 허가권 및 데이터 공유와 관련된 제약조건

4) 저장소와 백업

- 데이터의 백업 주기, 백업 위치, 복사본 수 (노트북, 컴퓨터 하드

디스크, 외장디바이스 저장을 지양)

- 데이터 보안 위협 요소 도출 및 처리방안
- 데이터 보안을 유지하기 위한 접근 제어 방법
- 협력자들의 안전한 데이터 접근 방법
- 야외에서 생산 및 수집되는 데이터의 안전한 전송 방법 등
- ISO 27001과 같은 공식적 표준 준수 여부

5) 데이터 선정과 보존

- 보존 혹은 파기 되어야 하는 데이터
- 유지 되어야하는 데이터 결정 방법
- 미래에 데이터를 사용할 예측 가능한 연구
- 데이터 보유 및 보존 기간
- 잠재적 재사용 가치 등
- 연구 발견 검증 및 새로운 연구 수행 또는 교육을 위해서 데이터가 어떻게 재사용될 수 있는지 기술
- 파일형식 변환 필요성 여부
- 데이터 보존비용

6) 데이터 공유

- 데이터 공유 대상과 공유 조건
- 데이터 공유채널로서 리포지터리 사용 여부

- 데이터 사용 요구의 직접 처리 여부
- 데이터 사용과 관련된 엠바고(embago) 기간
- 데이터 식별자 사용 여부
- 데이터 인용방법
- 데이터 공유와 관련된 제약조건
- 제약조건을 극복하고 최소화하기 위한 조치
- 데이터 배타적 사용권 기간과 이유
- 데이터 공유 협약의 필요여부 등

7) 책임과 자원

- DMP 시행, 검토, 수정 책임자
- 데이터 관리 활동 책임자
  - 데이터 획득
  - 메타데이터 생성
  - 데이터 품질 관리
  - 저장 및 백업 담당
  - 데이터 아카이빙 담당
  - 데이터 공유 담당
- 정책 준수 여부 확인자
- DMP 이행을 위한 필요자원 (전문가, 하드웨어, 소프트웨어)
- 데이터 리포지터리 이용을 위한 비용

본 지침서는 연구 데이터를 이해하고 관리하기 위한 전반적인 내용을 다루었다. 데이터 정의부터 연구 데이터 종류와 특징을 기술하였다. 마지막 장에서는 선진국에서 연구자들에게 요구되는 데이터 관리 계획서를 설명하였으며, 항목별 기술이 요구되는 내용을 정리하였다. 국내에서 데이터 관리 계획서가 연구현장에 요구되는 시점에 본 지침서가 널리 활용되기를 기대한다.

ISBN 978-294-0762-9

김선태, 이정훈, 정한민

연구데이터 이해와 관리

2017년 11월 00일 인쇄
2017년 11월 00일 발행

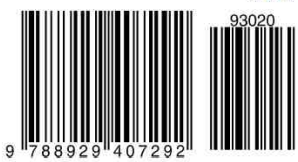
발 행 처

대전광역시 유성구 대학로 245
㉾ 34141
전화 : 042-869-1004
등록 : 1991년 2월 12일 제 5-259호

발행인
문 영 호

인쇄처
㈜

비매품



ISBN 978-89-294-0729-2