데이터 특성을 고려한 과학데이터 아카이브 시스템 설계를 위한 Data Curation Profile 분석

Analysis of Data Curation Profiles for Desinging a Science Data Archive System Considering Data Characteristics

임 종 태*, 서 인 덕**, 송 희 섭**, 유 승 훈*, 정 재 윤*, 조 중 권**, Aniruddha Paul**, 고 건 식**, 김 병 훈**, 박 윤 정**, 송 진 우*, 이 서 희**, 전 현 욱*, 최 민 웅**, 노 연 우*, 최 도 진*, 김 연 우*, 복 경 수*, 김 선 태***, 최 명 석***, 유 재 수*†

충북대학교 정보통신공학부*,
충북대학교 빅데이터학과**,
한국과학기술정보연구원 과학데이터연구센터***

Jongtae Lim*, Indeok Seo**, Heesub Song**, Seunghun Yoo*, Jaeyun Jeong*, Jungkwon Cho**, Aniruddha Paul**, Geonsik Ko**, Byounghoon Kim**, Yunjeong Park**, Jinwoo Song*, Seohee Lee**, Hyeonwook Jeon*, Minwoong Choi**, Yeonwoo Noh*, Dojin Choi*, Yeonwoo Kim*, Kyoungsoo Bok*,
Suntae Kim***, Myungseok Choi***, Jaesoo Yoo*†

School of Information and Communication Engineering, Chungbuk National University*
Department of Big Data, Chungbuk National University**
Scientific Data Research Center, Korea Institute of Science and Technology Information***

요약

본 논문에서는 과학 데이터 특성을 고려한 아카이브 시스템 설계를 위한 Data Curation Profile을 분석한다. Data Curation Profile은 생명과학데이터, 천문우주 데이터, 지질 데이터화학데이터, 사회과학데이터 등과 같은 13개의 과학데이터로 이루어져 있다. 13개의 과학데이터의 분석된 내용을 바탕으로 아카이브 시스템 설계시 고려해야 할 과학데이터 특성을 도출하고 이를 아카이브 요소 기술과 연계한다.

Ⅰ. 서론

최근 각종 실험 장비의 발전에 따라 천문, 항공, 우주, 유전자 등 첨단 분야에서 생산되는 과학데이터가 급격히 증대함에 따라 과학데이터에 빅 데이터 처리 기법의 적용이 요구되고 있다. 수집되는 과학데이터는 단순한 텍스트 데이터뿐만 아니라 음원, 이미지, 동영상 등 다양한 형태로 수집된다. 데이터 포맷뿐만 아니라 내부적으로 데이터는 단순 수치, 롱 시퀀스, 네트워크 그래프, 멀티미디어 등 특성에 따라 다양하게 분류할 수 있다. 따라서 데이터 포맷과 특성에 적합한 리포지터리 시스템 및 처리 기법의 선택이 매우 중요하게 요구되고 있다.

본 논문에서는 데이터 특성을 고려한 과학데이터 아카이브를 위한 Data Curation Profile 분석을 수행한다. 특성을 고려한 과학데이터 아카이브를 위한 아카이브 시스템 아키텍처 설계를 위하여 Data Curation Profile들을 분석한다. 분석된 내용을 바탕으로 과학데이터를 특성에 따라 분류하고, 분류에 따라 데이터 특성에 적합한 과학데이터 아카이브 시스템 아키텍처를 설계한다.

† 교신저자 : yjs@chungbuk.ac.kr

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2016-H8501-16-1013)과 2016년 KISTI의 '과학기술빅데이터 공동 활용 거버넌스 체제 지원' 과제의 재원을 지원받아 수행된 연구임

Ⅱ. Data Curation Profile 분석

본 논문에서는 데이터 특성을 고려한 과학데이터 아카이브를 위한 과학데이터 아카이브 시스템 아키텍처 설계를 위하여 Data Curation Profile[1] 분석을 수행한다. Data Curation Profile을 대상으로 해당 과학 분야에서 생성되는 데이터 크기, 데이터 포맷, 데이터 특성, 데이터 활용 측면으로 분석을 수행한다. 지면의 제한으로 분석한 13개의 과학데이터 중 대표적인 3개의 데이터를 중심으로 분석 내용을 제시한다.

1. Biochemistry

대표적인 생명공학 데이터로 Biochemistry 데이터가 존재한다. Biochemistry 데이터는 초파리의 뇌세포 데이터로부터 수집되는 데이터이다. 최초 실험 데이터는 단순 수치, 시퀀스, 이미지 데이터로 구성된다. 유전자 정보는 실시간 중합효소연쇄반응법(QCPR)을 통해 시퀀스 형태도 수집된다. 또한 유전자와 관련된 이미지는 .TIFF 포맷의 이미지 데이터로, 기타 관련 실험 데이터는 스프레드 시트 형태로 수집된다. 이렇게 수집된 데이터를 처

리하여 뇌세포간의 상호작용을 네트워크 그래프로 구성하며, 그래프는 스프레드 시트 형태로 구성된다. 처리된 데이터는 분석을 통해 그래프로 시각화하여 사용자에게 제공된다. 데이터의 크기는 38KB~6MB로 나타났다. 데이터 활용 측면에서는 데이터의 변경과 데이터간 연관성 정보가 일부 데이터에서 나타났다.

2. Astrophysics

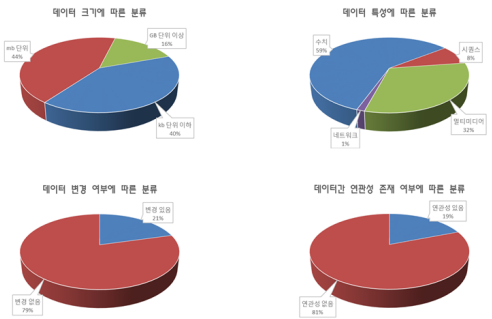
대표적인 천문우주 데이터로 Astrophysics 데이터가 존재한다. Astrophysics 데이터는 우주의 대규모 시뮬레이션을 통한 은사 성단과 그 특성에 관한 연구 데이터이다. 최초 실험 데이터는 HDF5 포맷의 다차원 배열의 데이터 집합이다. 관련된 입자들은 '스냅샷'이라는 하나의 군집으로 표현된다. 최종 결과는 3D 그래프로 시각화되어 사용자에게 제공된다. 데이터의 크기는 파일 하나당 30~70GB으로 나타났다. 데이터 활용 측면에서는 데이터의 변경은 존재하지 않았으며, 데이터의 연관성은 군집화된 스냅샷에서 일부 나타났다.

3. Soil Ecology

대표적인 지질 데이터로 Soil Ecology 데이터가 존재한다. Soil Ecology 데이터는 토양 특성에 영향을 미치는 다양한 생태계 성분을 분석하고, 토양 비옥도를 높이기 위한 다양한 치료에 대한 연구 데이터이다. Soil Ecology 데이터는 산성도, 이온화 지수, 토성, 유기물 등의 정보로 구성된다. Soil Ecology 데이터는 대부분이 단순 수치 데이터로 구성된다. 데이터의 크기는 5MB 이하의 매우 작은 용량으로 나타났다. 데이터 활용 측면에서는 데이터의 변경과 데이터간 연관성 정보가 나타나지 않았다.

Ⅲ. 결과 분석

본 논문에서는 생명공학, 천문우주, 지질, 화학, 인문사회 분야의 데이터를 분석했다. 그림 1은 Data Curation Profile 통계 분석 결과를 보여준다. 데이터 크기에 따른 분류에서는 KB, MB, GB단위의 데이터가 고르게 나타났다. 데이터 특성(구조 및 내용)에 따른 분류에서는 단순 수치 데이터와 이미지 데이터의 비중이 높게 나타났고, 시퀀스와 네트워크 데이터는 상대적으로 비중이 낮게 나타났다. 데이터 변경 여부와 연관성 존재 여부에 따른 분류에서는 약 20% 정도의 데이터들이 데이터 변경과 데이터간 연관성이 존재하는 것으로 나타났다. 분석 결과, 데이터의 크기, 데이터의 특성 그리고 데이터 변경 여부와 데이터간 연관성 존재 여부 와 같은 데이터의 활용 측면이 과학데이터를 분류하는데 의미 있는 특성임을 확인했다.



▶▶ 그림 1. Data Curation Profile 통계 분석 결과

그림 2는 Data Curation Profile 분석 내용을 바탕으로 도출한 데이터 특성을 보여준다. 데이터 특성은 데이터 크기, 데이터 포맷, 데이터 내용, 데이터 연관성, 데이터 접근성, 데이터 변경 여부, 데이터 사용 빈도 등으로 구분된다. 대표적인 데이터 특성으로 데이터 크기가 존재한다. 데이터 크기와 관련해서는 대용량의 데이터와 그렇지 않은(작은) 데이터로 구분할 수 있다. 대용량의 데이터의 경우, 분할하여 노드에 균등하게 저장 관리하기 위하여 분산 저장 관리, 부하 분산 관리 등의 기능이 요구된다. 반대로 작은 데이터의 경우 작은 파일에 대한 그룹화(병합, 카테고리화 등)가 요구되며 작은 데이터에 적합한 메타데이터 관리 등이 요구된다. 이렇게 도출된 데이터 특성은 데이터 특성에 기반한 과학데이터 아카이브 시스템 아키텍쳐 설계에 활용된다.

구분 항목	구분 기준	구분 의미	관련 아카이브 기술 요구사항
데이터 크기	큼	○	분산 저장 관리 부하 분산 관리 등
	중간	×	-
	작음	○	분산 저장 관리 스몰 파일 처리(병합) 메타데이터 관리 등
데이터 포맷	비구조적 파일	△	부분 검색 지원 등
	구조적 파일	△	부분 검색 지원 등
데이터 특성 (구조 및 내용)	수치, 텍스트	×	-
	시퀀스	△	부분 검색 지원 등
	네트워크 그래프	△	부분 검색 지원 등
데이터 연관성	있음	△	분산 저장 관리 데이터 캐싱 등
	없음	×	-
데이터 접근성	전체 접근	×	-
	부분 접근	△	부분 검색 지원 등
데이터 변경	Fixed 데이터	○	데이터 압축 저장 등
	Growing 데이터	○	변경 이력 관리 등
	Changeable 데이터	○	변경 이력 관리 등
데이터 사용 패턴	핫데이터 존재하지 않음	×	-
	핫데이터 존재	△	부하 분산 관리 데이터 캐싱 등

▶▶ 그림 2. 데이터 특성에 따른 아카이브 기술 도출 결과

Ⅳ. 결론

본 논문에서는 데이터 특성을 고려한 과학데이터 아카이브를 위한 Data Curation Profile 분석을 수행했다. 특성을 고려한 과학데이터 아카이브를 위한 아카이브 시스템 아키텍쳐 설계를 위하여 Data Curation Profile들을 분석했다. 분석된 내용을 기반으로 통계 분석을 수행하여 본 논문에서 고려한 데이터 특성들이 의미 있는 특성임을 확인하였으며, 데이터 특성에 따른 아카이브 기술을 도출했다. 추후 연구로는 데이터 특성에 따른 과학데이터 아카이브 시스템 아키텍쳐를 설계한다.

■ 참 고 문 헌 ■

[1] http://docs.lib.purdue.edu/dcp/