

클라우드 환경의 데이터 상호운용성 확보를 위한 표준 및 표준 활용 기술 동향

김장원 군산대학교 교수, ISO/IEC JTC1/SC32 위원

1. 머리말

1.1 클라우드 환경의 데이터 플랫폼 간 상호운용성 확보의 필요성

빅데이터 환경 속에서 다양한 형태와 대규모 데이터들의 생산이 급증하면서 데이터의 홍수라고 일컬을 정도로 방대한 양의 데이터들이 쏟아지고 있다.

이러한 빅데이터의 홍수 속에서 유의미한 데이터를 추출하는 것은 기업에서는 스마트한 고객 맞춤형 서비스를 제공하거나 정책적인 의사결정 지원 등과 같이 목적과 활용 방법 또한 다변화되고 있다. 그러므로 데이터의 체계적인 관리와 재활용에 대한 관심 또한 높아지고 있다.

특히, 분산 및 클라우드 환경에서 운영되는 서로 다른 시스템에서 생산된 데이터들의 연계와 통합은 데이터 활용을 통한 신규 비즈니스 창출에 이르는 데이터 활용의 전 주기적인 관점에서 주목받고 있다. 따라서 클라우드 환경에서 생산된 데이터의 상호운용성을 고려한 체계적이며 효율적인 관리 기술 및 방법이 필요하게 되었다.

최근 국내외에서는 다양한 데이터 서비스를 수행하는 데이터 허브, 데이터 포털 사이트 및 빅데이터 플랫폼 등이 개발되고 있으며 이러한 인프라를 통해서 각종 도메인 분야의 원시데이터뿐만 아니라 인공지능 모델에 활용을 위한 데이터가 생산 및 활용되고 있다.

그 결과 기존과 달리 다양하고 고품질 데이터의 생산 및 활용 등이 어려웠던 일반 시민 및 연구자들은 이러한 오픈 데이터로 다양한 인사이트를 도출하고자 점점 많은 관심을 가지고 데이터 생태계에 참여하고 있으며, 정부에서는 공공 및 오픈 데이터와 관련된 해커톤, 경진대회 개최, 산학연 지원 사업 등을 통해 데이터를 생산하고 데이터 허브 또는 데이터 플랫폼 등을 통해 공개하여 데이터에 대한 가치 확산에 노력을 기울이고 있다.

그 결과 수 많은 데이터들이 데이터 허브, 빅데이터 플랫폼, 포털 사이트 등을 통해 공개되고 활용하게 되었다. 그렇지만 특정 데이터의 경우 서로 고립된 형태로 공개되어 있어 데이터 활용을 위한 추가적인 전처리 작업을 통한 통합을 위한 노력이 필요한 경우가 발생한다. 또한, 사용자의 분석 목적에 따라 다양한 플랫폼에서 공개된 데이터들을 연계하고자 하는 경우 사용자들은 개별 플랫폼에 데이터에 대한 요청, 승인 및 접근을 통해 데이터를 획득하고 통합을 시도해 볼 수 있다.

예를 들어, 홍수로 인한 침수 피해 정보를 간접적으로 분석하길 원하는 사용자는 기상청의 동네예보 및 기상 관측 데이터, 행정 구역별 침수 데이터(지자체), 건축물정보(국가공간정보포털), CCTV 정보, 재난 피해액(공공데이터포털) 등의 데이터가 필요할 수 있다. 그렇지만 이러

한 데이터들이 서로 다른 플랫폼에서 제공되는 경우, 사용자들이 원하는 데이터들이 어떠한 플랫폼에서 어떠한 형태로 제공하는지 확인 과정이 필요하며 공개된 각각의 데이터들에 대한 메타데이터(항목명, 분류체계, 키워드), 인스턴스 검토를 수행하여야 해당 데이터들에 대한 통합 가능성을 판단할 수 있다.

앞의 예시처럼 사용자가 원하는 모든 데이터들이 오픈되어 사용자가 접근 가능하더라도 다음과 같은 이슈를 접할 수 있다. 일종의 기반 데이터(위치정보)를 이용하여 데이터를 통합할 때 위치정보(경위도 좌표)가 서로 다른 좌표체계로 표현되어 있을 수 있으며, 경위도 좌표값 들의 자릿수가 데이터마다 일치하지 않는 경우 공간 해상도가 상이할 수 있기 때문에 지리정보 기반 인스턴스 매핑을 위한 지오코딩과 같은 전처리 작업 등이 필요할 수 있다. 물론 위처럼 전처리 작업을 통해서 연계를 할 수 있는 데이터들이 있는 반면에 원시데이터 자체에서 제공하지 않는 일부 데이터(누락, 손실 등)의 경우 상호 연계 시 정보의 손실이 발생할 수 있다. 따라서 데이터 상호 연계를 위해 필요했던 소모적인 전처리 작업, 상호운용성 검토를 위한 수작업 및 추가작업, 정보의 연계 시 발생 가능한 정보 손실을 최소화할 수 있는 방법이 필요하다.

그러므로 데이터 플랫폼 간의 상호운용을 위한 방법은 데이터 간의 연계, 데이터 관리 체계 구축 등이 고려될 필요가 있으며 데이터 플랫폼 간의 상호운용성이 보장된다면 도메인/특정 목적을 위해 구축된 개별 데이터 플랫폼에서 제공하는 원시데이터의 가치를 높이고, 데이터의 재활용성을 확대할 수 있을 것이다.

1.2 데이터 상호운용성 확보 배경

클라우드 환경에서 운용되는 데이터 플랫폼들의 상호운용성을 보장하기 위해서는 플랫폼에서 제공하는 데이터 카탈로그, 데이터세트, 데이터 등에 대한 직접적인 연계 및 통합 방안이 필요하다.

데이터 카탈로그란, 원시데이터 그 자체의 정보뿐 아니라 다양한 원시데이터들의 연계 시 사용된 메타 정보를 포함하고 있다. 데이터 카탈로그는 일상생활에서 상품의 빠른 검색 시 도움을 주는 상품의 카탈로그와 그 목적이 유사한 것으로, 데이터에 대한 메타데이터를 정의하여 데이터 및 데이터세트에 대한 관리에 도움을 줄 수 있다. 데이터 카탈로그에서 정의할 수 있는 메타데이터는 카탈로그에 포함된 원시데이터 그 자체에 대한 정보뿐 아니라 원시데이터에 대한 데이터; 즉, 원시데이터의 생산 일자, 관리자, 공개자, 버전 등의 원시데이터의 관리를 위한 메타데이터를 포함할 수 있어 해당 데이터 카탈로그에 포함된 데이터와 데이터세트의 재활용 시 데이터에 대한 검색을 용이하게 할 수 있다. 그렇지만 이러한 데이터 카탈로그 또한 서로 다른 데이터 플랫폼에서 서로 다른 표현 방법을 통해서 공개하는 경우 데이터 카탈로그 수준의 상호운용을 위한 추가적인 방법이 필요하다. 따라서, 표준화된 방법으로 데이터 카탈로그를 정의하고, 카탈로그에 포함된 데이터세트를 명세하고 상호 연계를 방법이 필요하게 되었다. 이를 위해서 W3C는 DCAT(Data Catalog Vocabulary) 표준을 제정하였다. DCAT에 정의된 표준 어휘 및 개념을 이용하여 데이터 카탈로그의 정의, 카탈로그에 포함된 데이터 및 데이터세트에 대한 명세를 수행하여 데이터, 데이터세트, 카탈로그 간의 상호운용성을 확보할 수 있다.

2. 클라우드 환경의 데이터 표준 및 표준 활용 기술 동향

2.1 데이터 카탈로그 메타데이터 표준 소개(DCAT)

DCAT은 데이터 카탈로그들을 웹상에 공개하는 데이터 카탈로그들 간의 상호운용성을 보장하기 위해 사용할 수 있는 RDF 어휘들이 정의되어 있다. RDF (Resource Description Framework)는 W3C가 2004년에 공표한 표준으로 웹 자원들에 대한 메타데이터를 위한 데이터 모델과 함께 그래프 형태의 데이터에 대한 교환과 명세를 위한 방법들을 포함하고 있다. 따라서 RDF를 공통의 프레임워크로 사용하여 자원들이 포함하는 정보에 대한 의미의 손실이 없는 정보를 교환할 수 있게 하며 링크드 데이터(linked data)를 정의하기 위해서도 사용된다. DCAT은 버전 1.0이 2014년에 등장한 이후, 2022년 현재 DCAT 버전 3.0으로 진화해오면서 데이터 카탈로그를 표현하기 공통으로 사용 가능한 어휘뿐 아니라 데이터세트를 표현하기 위한 어휘, 데이터세트들을 관리하기 위한 속성, 웹 자원(resource)들 간의 관계 표현 방법 등의 명세를 포함하고 있다. DCAT은 추상적인 또는 원시의 데이터세트와 가공 및 배포되는 데이터 세트에 대한 구별을 위해 널리 사용되었다. 그 이후, DCAT 2.0은 다양한 도메인에서 사용되는 어휘들 간의 상호운용성을 지원하기 위해 보다 다양한 어휘들을 포함한 형태로 진화하였으며 이를 위해 7개의 클래스들과 속성이 추가되었다([그림 1]). dcat:Catalog 클래스는 데이터 세트 또는 데이터 서비스에 대한 메타데이터들의 집합으로, 한 개의 카탈로그에는 데이터 세트에 대한 메타데이터 정보들을 통해 자원들에 대한 명세가 가능하다. dcat:Resource는 데이터 세트, 데이터 서비스 또는 자원들에 대한 표현을 위한 것으로 dcat:Dataset의 상위(부모) 클래스이며, 특정 자원들에 대한 카탈로그를 보다 구체적으로 명세하기 위해 skos:Concept, dcat:Catalog 클래스와의 속성(관계)를 포함한다. dcat:DataService는 데이터세트와 서비스의 공개와 관련된 형태의 정보를 위한 클래스로 데이터세트 또는 서비스를 접근할 수 있는 속성(Access URL, API 등)을 포함하여 사용자들이 해당 데이터세트를 어떻게 접근해야 하는지를 표현한다.

이때 dcat:Distribution은 다운로드 대상의 파일에 대한 정보(접근 가능한 URL, 접근 형태)를 포함하고 있다. DCAT 3.0은 DCAT 2.0의 핵심 모델을 기반으로 일부 속성정보의 수정과 함께 신규 클래스 및 속성이 추가되었다. 예를 들어 dcat:Resource 속성을 통해서 dcat:Catalog와 dcat:Resource를 연결할 수 있게 되었으며 DCAT 2.0에서 dcterms:hasPart 속성을 대신하여 카탈로그와 자원들에 대한 관계를 자세히 표현할 수 있게 되었다. 더불어 자원(dcat:Resource)을 분류하기 위해 사용할 수 있는 테마(dcat:theme)를 OWL(Web Ontology Language)의 객체 속성(object property)처럼 사용할 수 있도록 하여, 자원에 대한 주요 분류체계를 특정 skos:Concept과 연결할 수 있으며 한 개의 자원은 다양한 테마를 가질 수 있게 되었다. 또한 원시데이터의 품질을 보장하기 위한 속성(spd:checksum, dcat:version, dcat:previousVersion, dcat:hasCurrentVersion)을 통해 데이터의 변형 유무와 버전 관리를 할 수 있다. 향후에는 연속적으로 연결되어 있는 데이터세트들을 관리하도록 dcat:DatasetSeries 클래스와 속성을 정의하여 데이터 생성 기준으로 데이터의 전후 관계를 표현할 예정(현재: 공표되지 않은 개발 중)에 있다.

세계 각국의 중앙정부 및 지방정부에서 인기를 끌기 시작하면서 2010년에 영국 정부에 의해 <http://data.gov.uk>에 도입되고 데이터를 개방할 수 있는 대표적인 플랫폼으로 그 가치를 인정받게 되면서 CKAN을 활용한 데이터 공유 생태계가 급성장하였다. 처음 40개의 데이터세트로 시작했던 <https://ckan.org/>은 점차 많은 데이터세트를 저장 및 관리할 수 있게 되었으며 2011년에는 Data Hub(데이터 허브)라는 새로운 브랜드로 명명되어 '데이터 허브'라는 개념이 전세계로 확산되는데 큰 기여를 하였다. CKAN은 오픈된 데이터를 찾는 카탈로그를 제공하는 기능에서 시작하였으나 현재는 정형, 비정형 등의 다양한 데이터 종류를 저장할 수 있는 저장소와 시각적인 분석을 지원하는 데이터스토어로서의 기능까지 제공하고 있다. CKAN과 DCAT은 이미 해외에서 검증된 사례가 많고, 정부에서 추진하고 있는 데이터 공유·유통 플랫폼 구축에 최적화된 '표준'과 '개방'을 지키는 플랫폼이다. 특히, CKAN은 OMI-PMH 기반 하베스팅(harvesting) 기능을 통해 해당 프로토콜을 포함한 CKAN 데이터 허브, GEOSS 데이터 허브 등에 저장된 데이터를 공유·연계할 수 있다. 지속가능한 데이터 생태계 조성을 위해서 CKAN에서는 데이터, 데이터 카탈로그에 대한 상호운용성을 지원하기 위한 다양한 확장 모듈을 개발 및 배포하고 있다. CKAN에서는 DCAT에서 정의한 어휘를 기반으로 데이터 카탈로그를 명세할 수 있으며 그 결과 표준화된 공통 어휘로 정의된 카탈로그에 대한 교환 및 연계가 가능하다.

2.2.2 CKAN에서의 DCAT 활용 방안

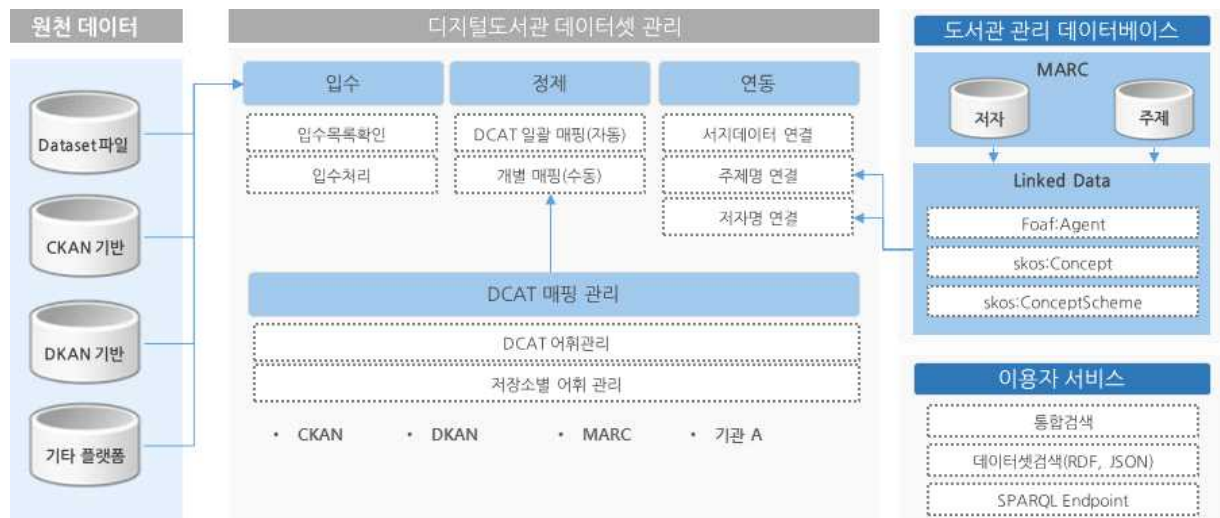
DCAT을 활용한 연구¹⁾에서는 새로운 지식정보자원으로 중요성이 높아지고 있는 최근의 데이터 생태계에서 데이터세트를 관리하고 서비스를 지원하기 위한 DCAT 활용 방법을 제안하였다. DCAT을 구성하는 8개의 클래스 중 핵심 클래스 4가지를 중심으로 클래스와 속성을 분석하고, 디지털도서관에서 DCAT을 기반으로 다양한 데이터세트를 관리하고 서비스할 수 있도록 한 모델을 제시하고 있다. 제안 모델을 통해서 원천데이터, 데이터세트의 관리, 링크드 데이터의 이용자 서비스로 구분하고 있다. 특히, 디지털도서관의 자원들을 정의하기 위해 기존에 정의하여 사용하던 다양한 메타데이터들의 상호운용성을 지원하기 위해 DCAT 매핑 기능을 제안하였다(그림 2).

이처럼 DCAT은 데이터세트의 관리를 위해 어휘, 데이터세트, 자원(resource)들의 관계 정보뿐 아니라 데이터의 분류체계와 분류체계를 구성하는 카테고리를 관리할 수 있도록 SKOS(Simple Knowledge Organization System)에서 정의한 `skos:ConceptScheme`과 `skos:Concept` 개념을 도입하여 각 분야에서 필요로 하는 도메인별 데이터 분류체계와 테마를 표현할 수 있다(<표 1>).

<표 1> skos:Concept을 이용한 어휘 사용 예

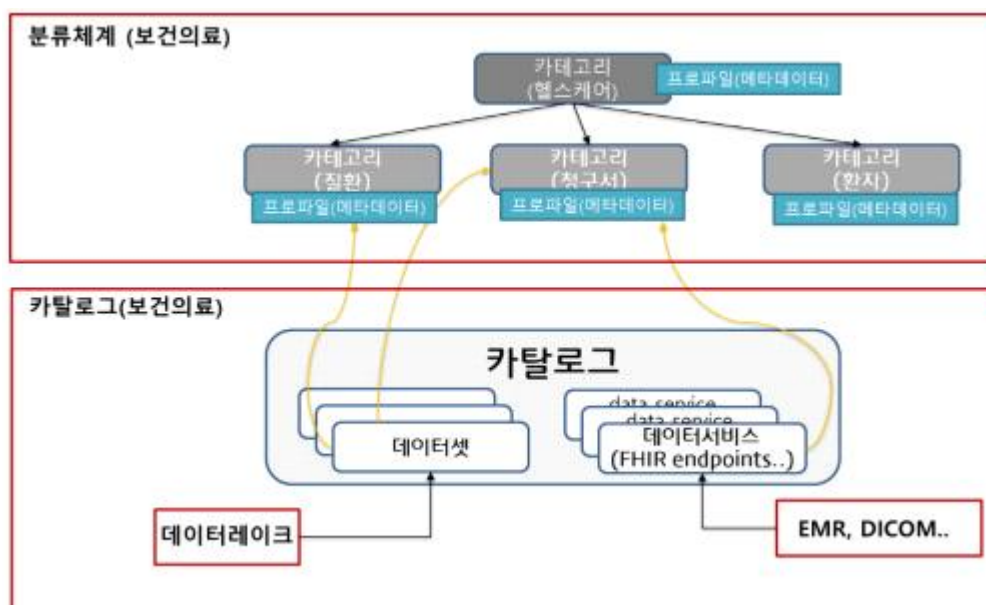
```
ex:animalThesaurus rdf:type skos:ConceptScheme;
    dct:title "Simple animal thesaurus";
    dct:creator ex:antoineisaac.
```

1) 박진호. (2019). DCAT을 활용한 디지털도서관 데이터세트 관리와 서비스 설계. 한국문헌정보학회지, 53(2), 247-266



[그림 2] 디지털도서관을 위한 DCAT 관리 서비스 시스템 구성도

따라서, DCAT 기반으로 정의된 데이터세트를 CKAN에 등록할 때 특정 도메인에 대한 카테고리를 선택할 수 있으며 기존에 정의된 분류체계를 통해 데이터 카탈로그를 매핑할 수 있다. 다음 그림은 보건의료 분야를 대상으로 분류체계와 카테고리 카탈로그 간의 관계 정의의 예를 보인다²⁾.



[그림 3] 데이터 분류체계 표현에 활용된 DCAT 예

2) 원희선. (2019). 데이터 유통·활용을 위한 W3C DCAT 표준 및 활용. OSIA Standards & Technology Review, 32(4), 19-22.

3. 국내외 DCAT, CKAN 적용 및 활용 사례 소개

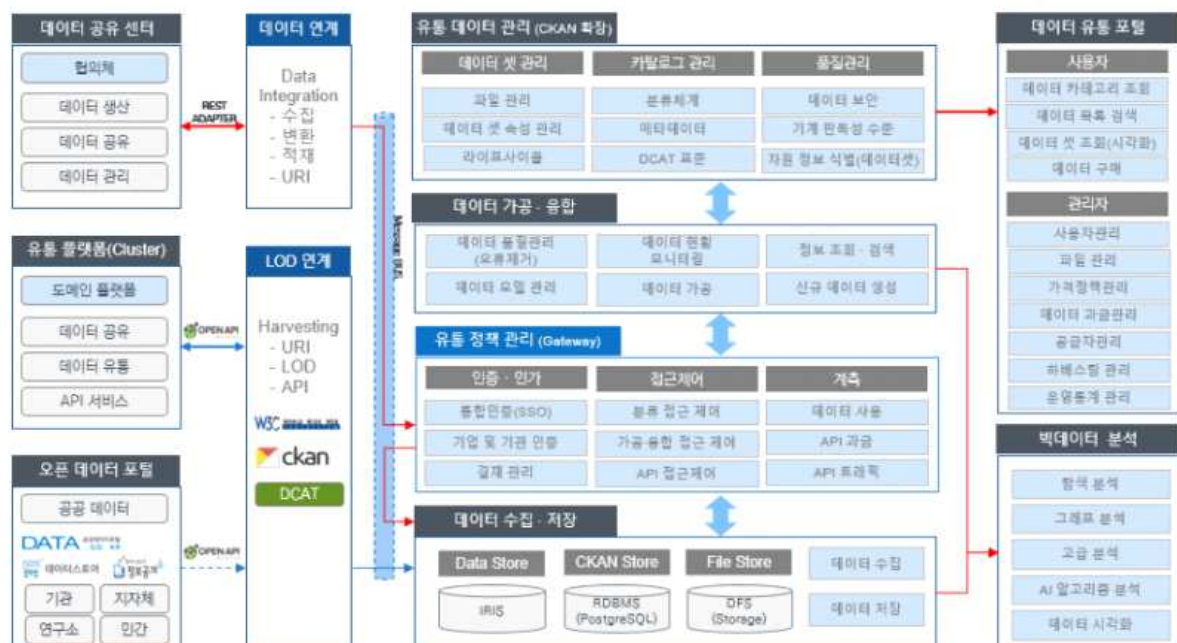
3.1 국내 CKAN 활용 사례

국내에서는 공공데이터 개방 포털(<https://data.go.kr/>), 서울시 정보 소통광장, 수원시 공공데이터 포털, 세종시 공공데이터 포털 등 국내 공공데이터 개방을 목적으로 도입하고 있다. 공공 목적으로 활용되는 사이트들은 포털 사이트와 마켓으로 구분될 수 있으며 포털사이트는 CKAN의 초기 도입 목표와 동일한 방향으로 데이터의 저장 및 카탈로그의 등록에 집중하고 있으며 마켓플레이스는 데이터에 대한 마켓 생태계를 구축하기 위한 용도로 활용되어 생산된 데이터들에 대한 가치에 따라서 유료 및 무료 형태로 판매 및 공유할 수 있다³⁾.

<표 2> CKAN을 활용한 데이터 플랫폼 현황 일부 예

구분	포털/마켓 구분	이름	운영 주체	서비스 개시 연도	플랫폼 정보	카탈로그	데이터 융합
공공기관	포털	공공 데이터 포털	행정자치부, NIA	2011	자체 플랫폼, CKAN	○	X
	마켓	데이터 스토어	한국 데이터 진흥원	2013	자체 플랫폼, CKAN	○	X
	포털	Data for Suwon	수원시	2015	CKAN API 제공	○	X
	마켓	빅데이터 허브	SKT	2013	자체 플랫폼	○	○
민간	마켓	오디피아	LG CNS	2016	자체 플랫폼	○	X
	마켓	API store	KT	2013	자체 플랫폼	○	X

국제표준 (DCAT) 기반의 데이터 관리 체계 및 데이터 마켓 플레이스 구축을 위한 통합 데이터 응용 플랫폼



[그림 4] DCAT, CKAN 적용한 마켓플레이스 개념도 (모비젠)

3) ㈜엔코아, CKAN을 활용한 민간 데이터 연계 활성화 방안 연구 결과 보고서, 한국데이터진흥원, 2017.10

3.2 해외 CKAN 활용 사례

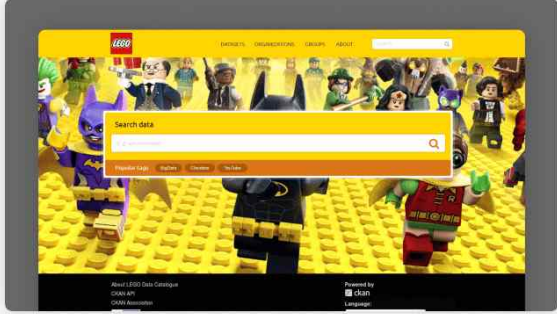
3.2.1 민간 분야

CKAN은 데이터 허브 역할 뿐 아니라 데이터 카탈로그를 통한 데이터의 상호 교환 지원, 데이터 분석 시각화 지원 등의 데이터 상호운용을 뒷받침하는 다양한 확장 기능들이 지속적으로 개발되고 있으며, 클라우드 환경에서의 CKAN 프레임워크를 운영하는 방법 및 데이터 공유 서비스를 제공하는 사례가 증가하고 있다.

대표적으로, LEGO Group은 글로벌 기업 데이터 공유 및 명세를 위한 카탈로그를 공유하기 위해서 CKAN을 도입하고 있으며, 전기 시스템 운영자 데이터 포털(ESO)은 영국의 에너지 데이터세트에 대한 허브 용도로 사용되고 있다([그림 5]). 그 밖에도 다양한 산업 분야에서 클라우드 환경을 통한 데이터 공유를 위한 플랫폼으로 CKAN을 선호하고 있다.

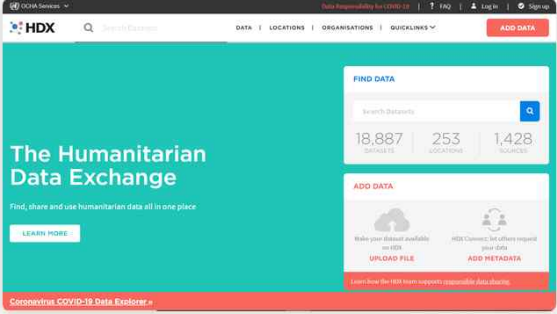
Powered by CKAN

Around the globe, enterprise organisations trust CKAN as their data management system of choice.



LEGO®

The office of the CIO for LEGO Group looked to CKAN to implement a proof of concept internal Data Hub as the basis for an internal catalogue for global corporate data holdings.



Humanitarian Data Exchange (HDX)

The Humanitarian Data Exchange uses CKAN as a feature that essentially allows organisations working with sensitive data sets to give visibility to their work without sharing the actual data.



Energi Data Service

Energi Data Service is a free and open data portal where anyone can get data about the Danish energy system such as CO2 emissions and consumption and production data.



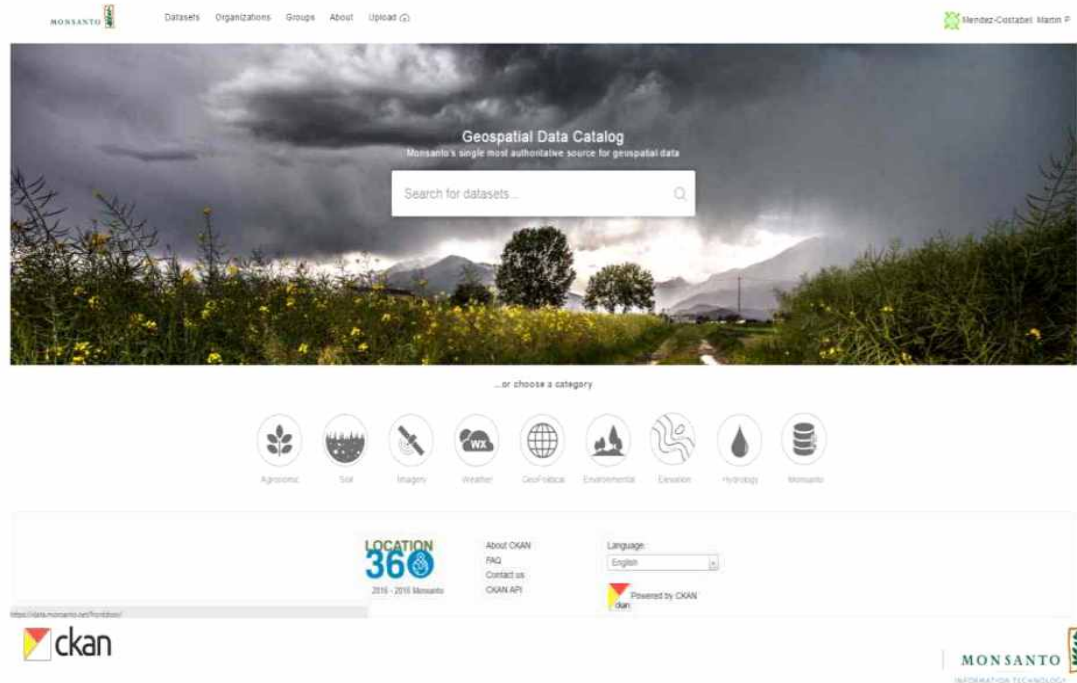
National Grid ESO Data Portal

The Electricity System Operator Data Portal (ESO) is one of the main custodians of energy data in Great Britain.

[그림 5] CKAN을 도입한 주요 기업(<https://ckan.org/enterprise>)

Link Digital사는 클라우드 서비스를 이용한 금융권 데이터를 CKAN을 통해 호스팅하고 있으며 농업 빅데이터 기업 몬산토는 클라우드에서 확장 가능한 CKAN 기반의 데이터 공유 플랫폼을 구축하고 기업 데이터, 고객 데이터, 공공 데이터를 연계 활용이 가능한 오픈소스 생태

계를 구축하였다([그림 6]). 이를 위해서 CKAN을 통한 카탈로그(데이터 목록) 공유 기술, 지리 정보의 각 특징(feature) 저장을 위한 PostGIS(데이터베이스), QGIS(지리정보시스템), AWS ECS(클라우드 호스팅서비스), amazon S3(클라우드 저장소) 등을 연계한 오픈플랫폼 기술을 적용하였다.



[그림 6] 농업 분야 지리정보 데이터 교환용 카탈로그 시스템 (몬산토)

3.2.2 공공분야 (연방, 지방정부)

[그림 7]은 CKAN 플랫폼을 도입한 정부 기관 목록의 일부를 보인다.



[그림 7] CKAN 플랫폼을 도입한 기관 목록 (일부)

■ 영국(Find open data – data.gov.uk, <https://www.data.gov.uk/>)

DATA.GOV.UK는 영국의 모든 공공기관 데이터 통합 서비스되는 데이터 포털로, CKAN을 사용하여 구축함⁴⁾

4) 안대진, & 이해영. (2013). 지방자치단체 데이터세트의 서비스 방안 연구 - 서울 열린 데이터 광장 서비스를 중심으로 -. 한국기록관리학회지, 13(2), 149-178, <https://doi.org/10.14404/JKSARM.2013.13.2.149>

■ EU (<http://publicdata.eu/>)

PublicData.eu는 범유럽 데이터세트 카탈로그를 제공하는 데이터 포털로 2011년 6월 포털의 백엔드가 CKAN으로 구성되어 유럽 지역 25개 정부 기관의 공공데이터 카탈로그를 수집할 수 있음

※ 수집 대상 기관 : 프랑스의 Data Publica, 스웨덴의 OpenGov, 이탈리아의 CSI Piemonte, 런던, 파리, 비엔나를 포함한 여러 시를 대상으로 함

■ 헬싱키 (<http://www.hri.fi/fi/>)

HRI(Helsinki Region InfoShare)는 국민들에게 지역 정보를 쉽고 빠르게 제공하는 것을 목표로 개발되었다. 기능과 데이터 카탈로그를 위해 CKAN을 핵심 코어로 사이트를 구축함

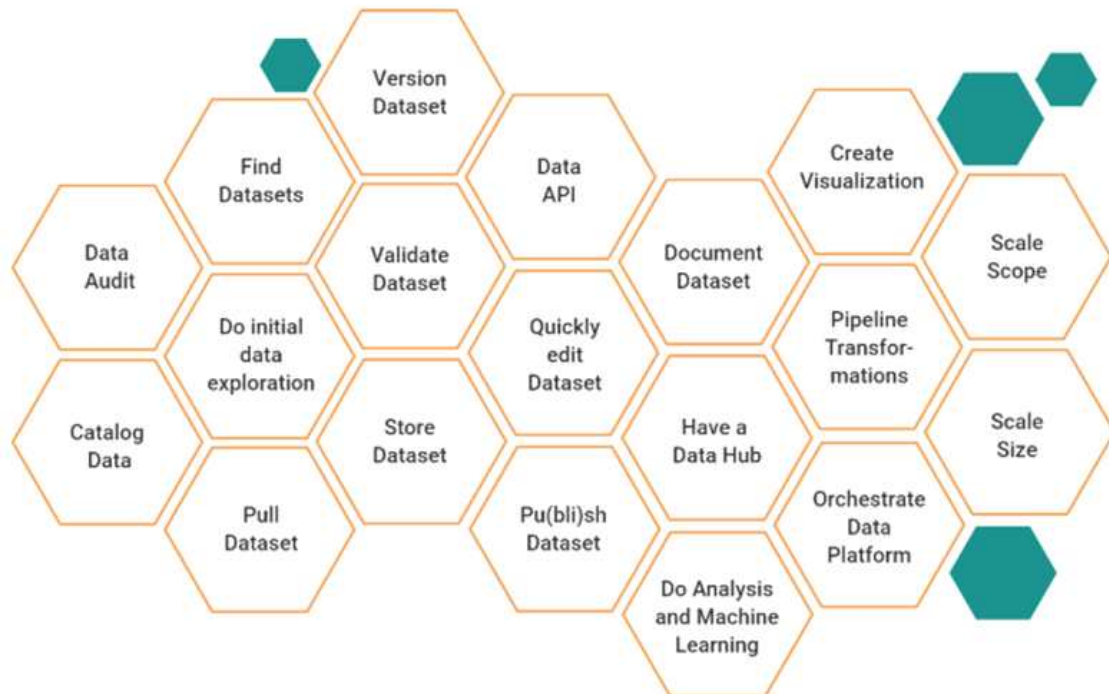
3.3 클라우드에서의 CKAN 활용

클라우드 환경은 분산된 곳의 자원을 통합하여 활용하거나, 대용량의 데이터세트 공유 시 하드웨어의 제약사항을 극복하기 위한 목적으로 선호도가 높다. 따라서 이러한 클라우드 환경의 특성을 고려하여 CKAN은 데이터 허브로서의 사용자의 데이터를 공유하기 위한 CKAN 클라우드, 또는 분산된 환경에 설치된 개별 CKAN을 효율적으로 관리하기 위한 관리자 관점에서의 서비스가 구축되고 있다.

DATOPIAN 사는 CKAN을 이용한 데이터 관리 시스템(Data Management System, 이하 DMS)을 위한 프레임워크를 제공한다. DMS는 데이터 포털, 데이터 카탈로그, 데이터 레이크와 같은 다양한 솔루션들을 통해 데이터 공유 플랫폼으로서의 역할을 수행함과 동시에, 다수의 CKAN을 클라우드 환경에서 운영에 필요한 CKAN 관리 및 운영 기술을 지원하고 있다. 이를 위해서 CKAN Cloud에 설치된 CKAN instance 관리를 위한 CKAN Cloud Operator 배포하여 AWS, Azure와 같은 클라우드 컴퓨팅 환경에서 CKAN을 효율적으로 사용할 수 있도록 지원하고 있다. 또한, 분산된 곳에 단편적으로 저장된 데이터, 대용량의 데이터들에 대한 통합을 위하여 데이터 파이프라이닝 작업과 다중 데이터 소스를 관리할 수 있는 클라우드 환경에서 지원한다⁵⁾. 또한, 최근 빅데이터 환경에서 사용되는 데이터는 그 유형, 크기, 특징이 다양하다. 이러한 특징을 포함한 데이터세트들을 저장하고 관리한다는 것은 원천데이터 그 자체를 보존해야 하는 것은 기본이며, 원천데이터를 설명하기 위해 정의된 메타데이터 정보 또한 지속적인 관리가 필요하다. 예를 들어 CKAN에 등록된 원천데이터에 포함된 내용 다수가 변경되어 버전이 변경된 데이터세트를 등록할 경우 기존 원천데이터에 대한 메타데이터를 일관성 있게 유지하면서도 신규데이터에 대한 메타데이터 또한 관리 대상으로 등록되어 관리해야한다. 특히, 시계열 데이터, 원천데이터의 가공에 따른 수정된 다양한 버전의 원천데이터 등이 그 주요 대상이며 이러한 원천데이터, 메타데이터, 카탈로그에 대한 품질 관리는 데이터의 통합 및 상호운용성 측면에서 중요하다. 그러므로 DATOPIAN에서는 데이터 관리 성숙 모델을 도입하여 CKAN에 저장된 데이터, 메타데이터, 카탈로그 등에 대한 수집, 통합, 상호운용성을 보장할 수 있는 서비스를 제공하고 있다. 이러한 데이터 관리 성숙 모델은 클라우드 환경에서의 데이터 거버넌스

⁵⁾ Datopian source hub, <https://github.com/datopian/ckan-cloud-operator>

체계를 견고하게 구축하기 위한 필수사항으로서 데이터들의 생성, 수집, 처리, 통합, 분석에 이르는 데이터 전주기 라이프사이클의 각 단계에 맞춘 데이터 품질 관리에 적용할 수 있다([그림 8]).



[그림 8] 클라우드 환경에서의 상호운용 지원을 위한 모듈 예 (DATOPIAN 사)

4. 맺음말

빅데이터, 클라우드 환경에 이르면서 데이터에 대한 관리는 대용량의 다양한 형태의 데이터 생산의 급증으로 인하여 일명 데이터 홍수의 시대가 도래하였다. 따라서 이러한 환경에서 데이터를 보다 효율적으로 사용하고 그 가치를 데이터의 라이프사이클 전반에 걸쳐 지속할 수 있도록 유지하기 위한 노력이 더욱 필요하게 되었다.

이러한 노력의 일환으로서 데이터세트에 대한 사용 목적별, 서비스, 도메인 등에 맞춘 카탈로그를 도입하여 데이터 관리 체계를 견고하게 만들고자 하였으며, CKAN이라는 플랫폼은 데이터와 관련된 정보를 저장하기 위한 저장소의 역할 뿐 아니라 카탈로그에 대한 관리 및 데이터의 상호운용성을 지원하는 기능까지 기술의 확장이 지속되고 있다.

그 결과 국내외의 민간 및 정부 기관 등에서 CKAN을 도입 및 적용하고 있으며 최근 클라우드 환경을 통한 CKAN 플랫폼의 운용 및 지원 사례 또한 증가하고 있으며 CKAN을 운영하기 위한 설치 및 서비스 환경에 대한 자유도가 확장되고 있다. 그렇지만 다양한 데이터의 정보를 손상시키지 않는 상호운용성을 보장하고 데이터세트에 대한 재활용성을 보다 높이기 위해서 다음과 같은 이슈가 여전히 남아 있다. 먼저 데이터세트에 대한 카탈로그를 정의함에 있어 전문 도메인 또는 국가별로 서로 다른 어휘, 분류체계 및 데이터를 식별하기 위한 식별체계가 상이할 수 있으므로 각 응용 분야 및 사용 목적에 맞는 확장이 필요하다. 본문에서 설명한 DCAT은 공통 어휘에 초점을 맞추고 있기 때문에 특정 응용 분야에 적합한 체계가 필요하다.

따라서 EU를 중심으로 DCAT-AP(Data Catalog Vocabulary-Application Profile)를 통해서 데이터 교환 시 응용 도메인과 관련된 어휘를 정의할 수 있다.

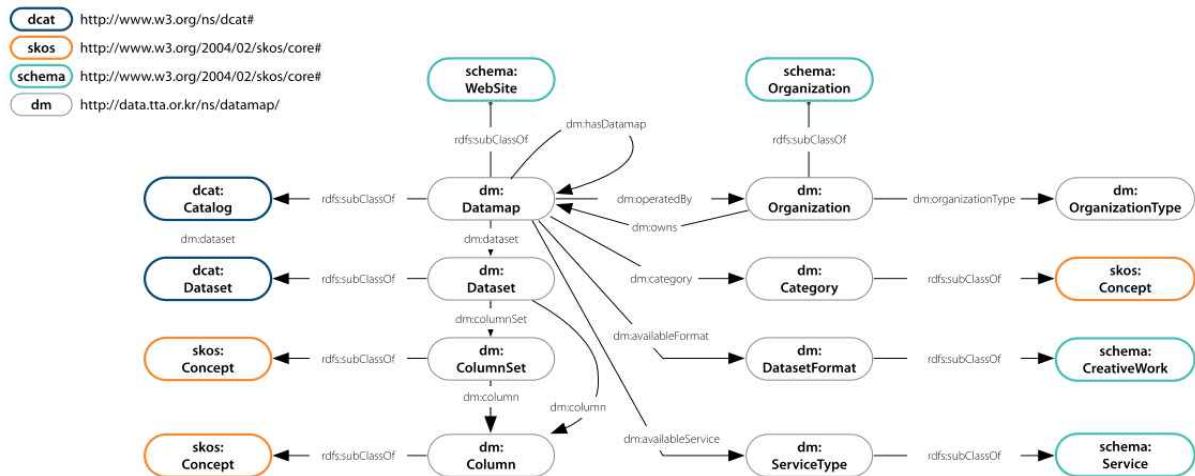
DCAT-AP는 다양한 데이터세트 유형들에 대한 메타데이터를 교환하기 위해 정의된 것으로 공통의 교환 형식을 통한 데이터 포털의 상호운용성을 지원하고 특정 커뮤니티 또는 도메인에서 공통으로 사용하는 응용 프로파일(AP)을 확장할 수 있도록 하여 도메인 데이터에 적합한 메타데이터 요소들을 명세할 수 있다. 대표적인 AP에 대한 응용 사례로, StatDCAT-AP는 EU 데이터 포털 사이트에서의 통계 정보를 담고 있는 데이터세트, 데이터시리즈 및 서비스를 명세하기 위한 전문 어휘를 정의하고 있다. 그러므로 오픈된 형태의 다양한 공공데이터 포털에 존재하는 통계 데이터를 보다 효과적으로 통합할 수 있으며 통계 데이터 검색을 용이하게 한다.

<표 3> StatDCAT-AP에서 사용하는 대표 어휘 예(stat:numSeries)

Property	Number of data series
URI	stat:numSeries
Subproperty of	dct:extent
Range	dct:SizeOrDuration, expressed as xsd:integer
Definition	The actual number of series in the data set as referenced in the Distribution.
Comment	The Cartesian Product is the number of modalities of each dimension, excluding what Data Cube calls the measure dimension (that denotes which particular measure is being conveyed by the observation). This is the total of the theoretical number of combinations of dimension values. The actual number of series (as represented in numSeries) might be less than the theoretical number calculated as the Cartesian Product, and when combined with the dimension list, a useful indication of the detail of the data in the data set)often called the 'density'.

따라서 응용 프로파일링이 사용되는 도메인 및 서비스에 맞춘 어휘와 함께 어휘를 체계적으로 정의할 수 있는 분류체계가 확장될 필요가 있으며 이를 통해 다양한 도메인 데이터의 상호운용을 지원할 수 있으며 데이터세트에 대한 검색의 정확성과 재현율을 높일 수 있다.

현재 국내에서 활용 가능한 어휘에 대한 표준을 위한 DCAT-AP-KR이 정보통신단체표준화위원회(TTA)의 빅데이터 프로젝트 그룹(PG1004)에서 과제로 채택되어 표준화를 추진하고 있다. 또한, 동일 프로젝트 그룹에서 2021년 제정된 데이터 맵 어휘 단체 표준은 데이터 서비스에서 사용되는 메타데이터를 기계가 읽을 수 있으며 활용하기 위한 목적으로 제정되었으며 데이터 맵 어휘를 통해서 다양한 포털에 등록된 데이터에 대한 검색이 가능한 것을 보인다([그림 9]). 이를 통해서 데이터를 서비스하는 주체는 보유 및 제공하고 있는 데이터의 목록, 데이터의 유형 등과 같은 관리적인 측면과 함께 기초 통계 정보(조회, 다운로드 회수)를 함께 제공할 수 있다. 따라서 정부와 민간에서 공개된 데이터 포털이 데이터 서비스를 제공함에 있어 표준화된 방법으로 데이터 맵 어휘를 적용할 경우 정보를 교환하고 통합하는데 활용할 수 있다.



[그림 9] 데이터맵 어휘의 핵심 클래스와 속성

또한, 데이터 통합과 상호운용성 보장을 위해서 데이터 허브에 대한 운영 전략 또는 거버넌스 체계 도입이 선행될 필요가 있다. 예를 들어, 앞서 설명한 EU 데이터 포털의 경우 모든 데이터를 하나의 포털에서 보유하는 전략 대신 각 기관의 데이터 포털에서 개별 데이터를 유지하고 데이터에 대한 카탈로그만을 EU 데이터 포털에 유지하는 방법을 채택하고 있다. 이러한 전략은 각 기관 별 수립된 거버넌스 체계를 통해 원시데이터의 라이프사이클 관리 및 메타데이터 표준화를 수행함으로써 기관에서는 데이터에 대한 오너십을 획득하고 데이터 통합시 관리에 대한 책임을 분담할 수 있다. 또는, 앞서 대표 사례로 설명한 DATOPIAN사에서 운영하는 것처럼 개별 CKAN 인스턴스 또한 클라우드 환경에서 운영할 수 있도록 하고, 그러한 CKAN에 대한 전사적인 통합까지도 수행할 수 있게 된 것을 알 수 있다. 다시 말해서 대용량의 데이터 저장에 대한 효율성과 중복된 데이터들에 대한 저장 공간의 낭비 등을 예방하기 위한 용도의 클라우드 환경의 통합은 클라우드 컴퓨팅 환경이 우수한 현재의 시점에서 도입할 수 있는 전략으로 도입할 수 있을 것이다. 특히 클라우드 환경을 통한 통합 환경을 구축할 경우 각각의 인스턴스 담당을 위한 업무, 체계적인 포털 시스템 운영을 위한 통합적인 가이드라인에 대한 노력이 개별 운영환경보다 많이 줄어들 수 있을 것이다.

[참고 문헌]

- [1] 데이터맵 어휘, TTA.KO-10.1291, 2021.
- [2] 데이터맵 어휘 표준 소개, TTA 저널, 2021.
- [3] 데이터 카탈로그 어휘, TTA.E.O-10.0427, 2017.
- [4] 데이터 카탈로그 어휘 V2, TTA.E.O-10.0427/R1, 2020.
- [5] W3C, DCAT 3.0, <https://www.w3.org/TR/vocab-dcat-3/> (access date, 2022.8)
- [6] W3C, SKOS, [<https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/>] (<https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/>) (access date, 2022.8)
- [7] EU 데이터 포털, <http://publicdata.eu/>, (access date, 2022.8)
- [8] 영국 데이터 포털, <https://data.gov.uk/>, (access date, 2022.8)

- [9] 헬싱키 데이터 포털, <http://www.hri.fi/fi/>, (access date, 2022.8)
- [10] 우크라이나 데이터 포털, <https://data.europa.eu/en> (access date, 2022.8)
- [11] StatDCAT-AP 1.0.0, ISA working group. "StatDCAT-AP-DCAT Application Profile for description of statistical datasets.", 2016.
- [12] CKAN을 활용한 민간데이터 연계 활성화 방안 연구 결과 보고서, 한국데이터진흥원, 2017
- [13] Winn, Joss, "Open Data and the Academy: An Evaluation of CKAN for Research Data Management", 2013
- [14] 빅데이터 유통을 위한 카탈로그 메타데이터 표준 동향 및 서비스 구조, 안양대학교, 2019
- [15] DCAT Application profile for data portals in europe. ISA Programme of the European Commission, 2015
- [16] 박기성, 김태연, 누를아민, 서호진, 김현욱, 원희선 and 이영구. "CKAN을 위한 DCAT-AP 기반의 프레임워크의 설계 및 구현." 데이터베이스 연구 33.3 (2017): 40-51.
- [17] 영국, EU, 헬싱키 출처 - 안대진, & 이해영. (2013). 지방자치단체 데이터세트의 서비스 방안 연구 - 서울 열린 데이터 광장 서비스를 중심으로 -. *한국기록관리학회지* , 13(2), 149-178, <https://doi.org/10.14404/JKSARM.2013.13.2.149>
- [18] Naghman Waheed, Martin Mendez-Costabel (2017.03.15.), "Enterprise Geospatial Platform – A perfect fusion of cloud and open source technologies", MONSANTO, Strata + Hadoop World. (access date, 2022)
- [19] CKAN 공식 홈페이지, <https://ckan.org> (access date, 2022)
- [20] CKAN-DCAT Github, <https://github.com/ckan/ckanext-dcat>
- [21] DATOPIAN CKAN Cloud Github, <https://github.com/datopian/ckan-cloud-operator>
- [22] CKAN Github, <https://github.com/ckan/ckan>
- [23] 아마존 CKAN 마켓플레이스, <https://aws.amazon.com/marketplace/pp/prodview-ag6z2fqishchg> (access date, 2022.8)

※ 본고는 한국정보통신기술협회에서 수행 중인 '클라우드 상호운용성 확보' 사업에서 관련 기술 전파를 위해 의뢰한 보고서로(22년도 제2호), 클라우드 환경에서 데이터 상호운용성 확보를 위한 표준 및 표준을 활용한 기술 동향에 대해 소개하고자 하는 목적으로 작성됨.