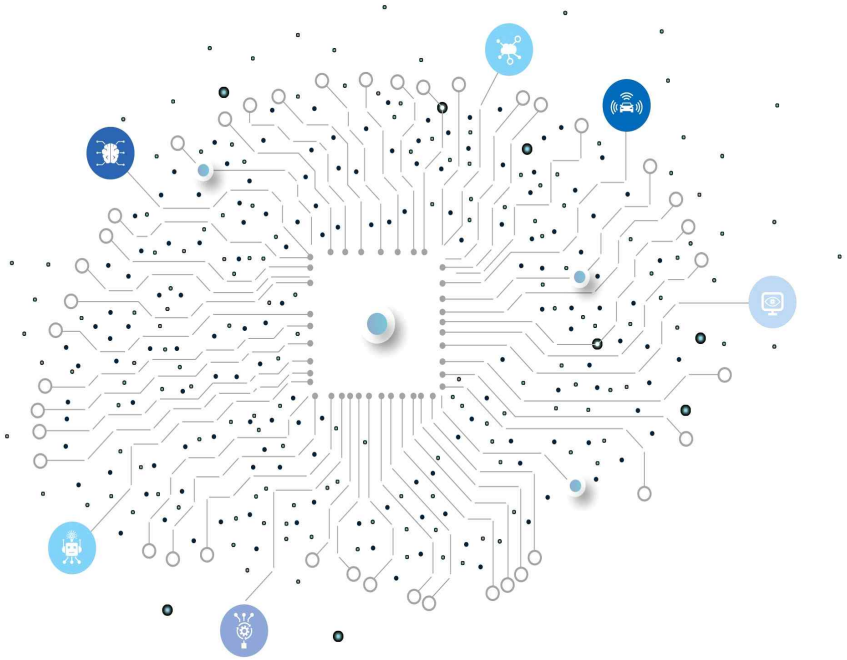
AI Network Lab INSIGHT Vol.5 (2021.05)



데이터 유통 및 데이터 품질 기술동향

| 작 성 | 안양대학교 정의현 (jung@gs.anyang.ac.kr)

발 행 처 한국지능정보사회진흥원

발 행 인 문용식

기 획 한국지능정보사회진흥원 지능형인프라본부 미래네트워크센터

보고서 온라인 서비스 www.nia.or.kr

NIA 한국지능정보사회진흥원 NATIONAL INFORMATION SOCIETY AGENCY

Contents

보고서 주요 내용

개요

○ 인공지능이 전 산업 영역으로 빠르게 확산되고 이를 통한 사회변화가 가속화되는 상황에서 인공지능 성능의 토대가 되는 양질의 데이터 확보는 4차 산업혁명 산업 발전을 위해서는 무엇보다 시급한 문제가 되었다. 이러한 상황에서 데이터 유통을 촉진하는 데이터 카탈로그 메타데이터의 이해와 적용은 매우 중요하다.

○ 데이터 카탈로그의 실현을 위해 W3C는 2014년에 DCAT(Digital Catalogue Vocabulary)을 제시하였고, 2020년에는 DCAT 2.0 버전으로 개정하였다. EU도 DCAT을 확장한 DCAT-AP(DCAT Application Profile) 규약을 2015년에 제정하였고, 2020년에 2.0.1 버전으로 개정하였다.

○ 데이터의 품질을 평가하기 위해서도 메타데이터는 매우 중요한데, 데이터 자체로는 품질을 평가할 수 없기 때문에 대상 데이터에 최신성, 정확성, 상호연계성 등의 지표를 측정하여 메타데이터로 추가함으로써 데이터의 가치를 산정할 수 있다. DQV와 같은 표준적인 방식으로 데이터 품질이 평가되고 가치가 산정되면, 데이터에 적절한 가격이 매겨질 수 있고 데이터 생산자의 생산 의욕과 데이터 유통의 신뢰도를 높일 수 있게 된다.

○ 세계 각국은 데이터 생성과 유통을 가속화하고자 하는 노력을 하고 있으며, EU 데이터 포털의 사례에서도 볼 수 있듯이 데이터 유통과 데이터 품질 표준을 적극 활용하고 있다. 국내에서도 한국지능정보사회진흥원이 추진하는 통합데이터 지도(https://www.bigdata-map.kr) 사이트가 DCAT 2.0을 기반으로 구축되고 있으며, 품질 부분이 보완된다면 더욱 활성화될 것으로 기대된다.

주 요 내 용

1. 데이터 생태계를 위한 메타데이터의 중요성

인공지능 기술 경쟁력이 국가 경쟁력으로 간주되고 있는 4차 산업혁명 환경에서 데이터는 혁신적인 가치를 창출하는 새로운 자원으로 주목받고 있으며, 방대한 양의 데이터에서 의미있는 정보를 추출하고 경제적 가치를 창출하는 빅데이터와 인공지능 기술은 4차 산업혁명의 핵심 기반 기술로서 산업 전반에 큰 파급효과를 불러올 것으로 예상되고 있다. 주요 선진국들은 데이터 유통 생태계가 4차 산업혁명 기술의 핵심이라는 것에 주목하여 데이터를 효율적으로 유통하기 위한 다양한 방안을 제시하고 있다. 영국의 "오픈 데이터 정책"과 같은 것이 이러한 노력의 일환이며, 특히 EU의 경우 다양한 국가가 하나의 연합체에 들어 있기 때문에 개별 국가에서 산출된 데이터 공유나 유통에 대한 노력이 일찍부터 진행되어 왔다. 이러한 노력은 EU Open Data Portal과 European Data Portal의 발족과 이 두 데이터 포털이 통합된 EU의 공식 데이터포털 (https://data.europa.eu/)을 기반으로 한 생태계 구축으로 이어졌고, 2021년 5월 현재 이 데이터 포털을 통해 EU에서 생산된 130만개 이상의 데이터가 유통되고 있다.

이렇듯 데이터의 유통을 통한 데이터 생태계 활성화를 위해서는 데이터 포털이 매우 중요하나, 아쉽게도 데이터 포털의 구축만으로는 생태계 활성화가 되지 않는다는 점이 오랫동안 지적되어 왔다. 기본적으로 데이터 포털에 탑재된 데이터는 사이즈가 크기 때문에 소비자 측에서 데이터의 내용을 일일이 살펴볼 수 없다. 또한 생산자 입장에서도 복사가 구매 행위의 완성인 디지털 경제에서 데이터를 모두 제공해서는 수익을 기대할 수 없다. 이러한 이유 때문에 데이터 유통 생태계에서는 개별 데이터의 특성을 제시하는 메타데이터(metadata)의 도입은 필연적이다. 일반적으로 데이터를 위한 메타데이터는 기존의 다양한 메타데이터 유형 및 비정형 데이터를 위한 최소한의 코어 메타데이터를 제공함으로써, 이들을 통합, 연계, 관리할 수 있어야 한다.

메타데이터(Metadata)는 "데이터에 관한 데이터"라는 의미이다. 예를 들어 도서관에서 서적을 검색하는 경우, 필요한 서적을 찾기 위해 모든 서적의 내용을 전부 조사하는 것은 무척 힘들고 소모적이므로, 서명, 저자명과 같은 "서지데이터"를 이용하게 된다. 이런 서지데이터처럼 책의 내용(데이터)을 추상화하여 "해당 데이터에 대한 데이터"로 정리하는 것이 메타데이터이다. 즉, 데이터의 대략적인 파악을 위해서 데이터 자체의 내용을 모두 살펴보기 보다는 데이터에 붙은 꼬리표인 메타데이터를 살펴보면 된다. 물론 메타데이터 작성과 유지에는 수고와 노력, 비용이 필요하게 된다. 하지만, 일단 정리된 메타데이터는 데이터 검색과 유지 보수에 막강한 위력을 발휘하게 된다.

데이터의 품질을 평가하기 위해서도 메타데이터는 매우 중요한데, 데이터 자체로는 품질을 평가할 수 없기 때문에 대상 데이터에 최신성, 정확성, 상호연계성 등의 지표를 측정하여 메타데이터로 추가함으로써 데이터의 가치를 산정할 수 있다. 표준적인 방식으로 데이터 품질이 평가되고 가치가 산정되면, 적절한 가격이 매겨질 수 있고 데이터 생산자의 생산 의욕과 데이터 유통의 신뢰도를 높일 수 있게 된다. 따라서 데이터 품질관리에 관한 연구는 경영관리 사이클(Plan-Do-See) 관점에서 재해석되고 있으며, 학계 및 산업계에서 데이터 품질에 대한 중요성을 새롭게 인식하고 있다. 그리고 이러한 데이터 품질을 기술하기 위한 데이터 품질에 관한 메타데이터 표준과 정책이 제안되고 있다.

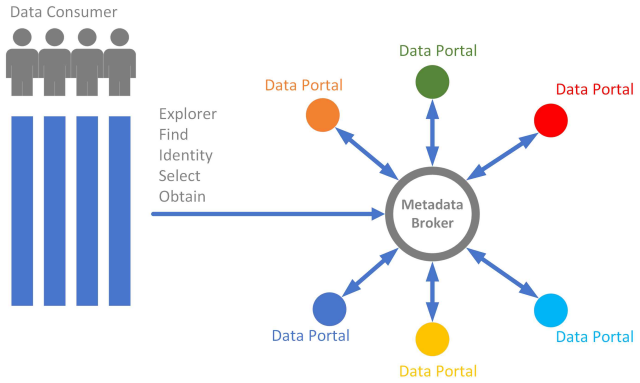
2. DCAT 표준

(1) 데이터 카탈로그 개요

메타데이터의 도입은 데이터 생태계 활성화 측면에서 권장되어야 함에도 불구하고, 거래소 별로 독자적인 메타데이터를 도입하는 것은 메타데이터 호환성이 떨어지기 때문에 전체 생태계 측면에서 바람직하지 못하다. 즉, 데이터 포털이 많아지면 많아질수록, 메타데이터의 호환성 문제로 인한 파편화 현상이 발생하며, 이로 인해 데이터 포털이 많아지면 원하는 데이터를 찾기가 더 힘들어지는 모순적 상황에 직면하게 되기 때문이다. 물론 이러한 메타데이터를 이용한 유통 구조는 그 방법과 예상 결과물이 매우 우수함에도 불구하고, 거래소마다 각자의 비즈니스 전략과 데이터 관리 방식이 상이하기 때문에 하나의 메타데이터 형식을 강제하는 것은 현실적으로 쉽지 않은 문제이다.

데이터 포털의 난립으로 인한 파편화 현상을 해결하기 위하여, 개별 거래소들의 고유 메

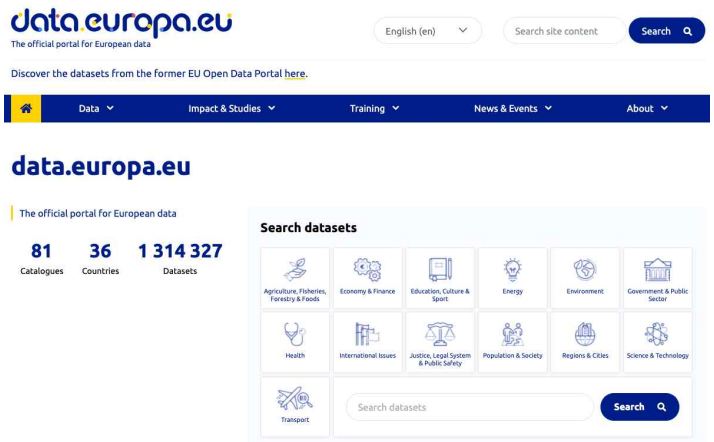
타데이터의 호환을 위한 메타-메타 데이터(Meta-metadata)의 필요성이 생겼고, 이러한 목적으로 데이터 카탈로그(Data Catalog)라는 개념이 EU와 W3C를 중심으로 제안되었다. 데이터 카탈로그란 "특정 데이터 포털이 어떤 종류의 데이터를 갖고 있고, 데이터가 어떤 형태를 갖고 있는 지를 표시하는 말 그대로 카탈로그 역할을 하는 메타데이터이다. 이를 이용하면, [그림 1]과 같이 데이터 포털의 독립성은 최대한 보장하면서, 데이터 포털이 보유한 데이터를 데이터 카탈로그 메타데이터로 제공하여, 이를 통한 데이터 검색과 유통이 가능한 상호운용 구조를 만들 수 있게 된다.



[그림 1] 데이터 카탈로그 메타데이터를 이용한 데이터 상호운용 구조

데이터 카탈로그의 실현을 위해 W3C는 2014년에 DCAT(Digital Catalogue Vocabulary)을 제시하였고, 2020년에는 DCAT 2.0 버전으로 개정하였다. EU도 DCAT을 확장한 DCAT-AP(DCAT Application Profile) 규약을 2015년에 제정하였고, 2020년에 2.0.1 버전으로 개정하였다. 이 두 표준은 현재 EU에서 채택되어 EU 데이터 포털(https://data.europa.eu)의 기본 메타데이터로 사용되고 있다. DCAT은 시맨틱웹 기술인 RDF(Resource Description Framework)를 이용하여 데이터 포털에 저장된 데이터 셋에 대해 기술할 수 있는 문서 기술 방식이다. 이 표준들의 특징은 개별 데이터 포털에서 생성되는 데이터의 메타데이터는 존중하면서, DCAT이나 DCAT-AP로 메타데이터 카탈로그를 만들어 해당 데이터들의 호환을 제공한다는 점이다. 이러한 접근 방안은 기존 데이터 제공자들이 이미 투자한 정보 자산을 포기하지 않고도 데이터 유통 생태계로 편입되는 효과를 낳게 하였다. 특히, DCAT-AP의 경우에는 분야별 AP의

정의를 지원하기 때문에 통계, 지리정보, 대중교통 등에서 이미 개별 AP가 정의되어 있다. EU는 여러 국가의 연합체이기 때문에 2014년에만 160개 이상의 데이터 포털이 존재하고 있으며, 이의 호환이 중요한 문제로 대두되었다. EU는 이를 해결하기 위하여 [그림 2]와 같은 EU 데이터 포탈(https://data.europa.eu/)을 제시하였다.

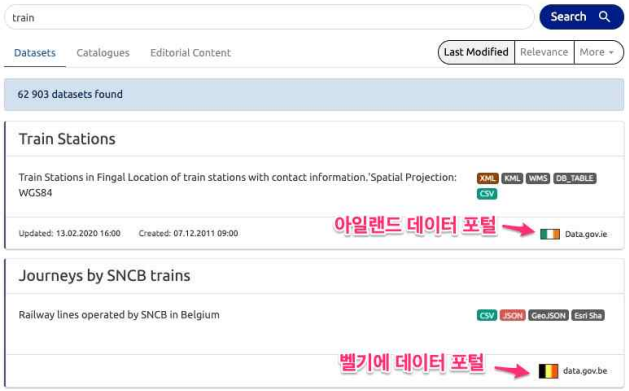


[그림 2] EU 데이터 포털

EU 데이터 포털은 EU의 모든 데이터를 하나의 포털에서 유지하는 전략을 택하지 않고, 개별 데이터는 해당 데이터 포털이 유지하고 대신에 데이터에 대한 카탈로그만을 EU 데이터 포털에 유지하는 방법을 채택하였다. 예를 들어, [그림 3]과 같이 train라는 키워드 검색을 하게 되면, 간단한 경우에는 EU 데이터 포털에 저장된 train 데이터를 보여주기도 하지만, 기본적으로 train라는 키워드를 포함하는 개별 데이터 포털에 대한 정보와 링크를 보여준다.

[그림 3]의 검색 결과 화면에서 아일랜드 포털의 링크를 선택하면, 해당 데이터에 대한 메타데이터를 이용해 실제 데이터를 다운로드 받을 수 있는 링크와 부가정보가 표시된다. 즉, 실제 데이터는 EU 데이터 포털이 아니고, 해당 데이터 포털인 아일랜드 데이터 포털에서 받을 수 있도록 구성되어 있다. 이렇게 데이터 카탈로그를 통한 검색이 가능한 이유는 개별 데이터 포털에서 DCAT 기반의 데이터 카탈로그를 제공하기 때문이다. 이렇듯, EU에 많은 데이터 포털들이 있고 해당 데이터 포털들마다 독자적인 데이터 보유 방식과 관리 정책을 갖고 있지만, 데이터 카탈로그를 제공함으로써 데이터 소비자들이 쉽게 데이터를 검색하고

활용할 수 있게 해준다.

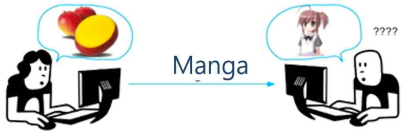


[그림 3] EU 데이터 포털 검색 결과

(2) RDF 표준

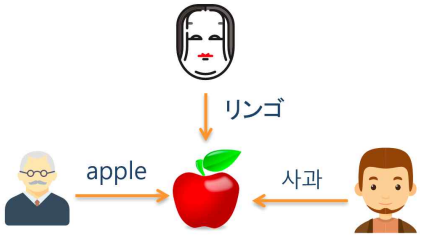
DCAT 표준은 시맨틱웹 기술이 바탕인 RDF 표준으로 구성되어 있기 때문에, RDF 표준을 이해하지 못하면 DCAT 표준을 제대로 이해할 수 없다. 따라서 본 고에서는 RDF 표준에 대해서 간략히 설명하고자 한다. 팀 버너스리에 의해 만들어진 웹(World Wide Web)은 인간이 인터넷에 흩어진 정보를 시각적으로 파악하는데 커다란 기여를 하였다. 그러나 웹이 시각적 표현에 집중하면서, 컴퓨터가 이해할 수 있는, 혹은 동일한 시맨틱을 제공하는 정보 구조의 필요성이 대두되었다. 이를 해결하기 위하여 1994년에 팀 버너스리의 제안으로 시맨틱 웹에 대한 연구가 W3C를 중심으로 심도 있게 논의되었다.

시맨틱 웹은 표현(presentation)에 집중한 기존 웹과 달리 개별 정보를 노드로 정의하고, 정보와 정보를 링크로 연결하는 의미망(Semantic Network) 구조로 모든 정보에 URI를 할당해서 유일성을 보장하는 특징을 갖고 있다. 예를 들어, [그림 4]에서 볼 수 있는 것처럼 기존 정보는 표층적 레벨(syntactic level)에서 단일한 표현을 갖고 있다. manga라는 표현은 포르루칼어로는 망고 열매를 의미하지만, 많은 나라에서는 일본 만화의 영어식 표현으로 이해하게 된다.



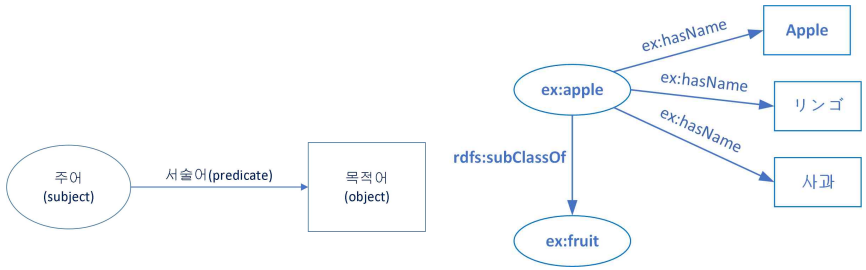
[그림 4] 의미의 표층적 표현

이에 비해 [그림 5]의 경우에는 "사과"라는 의미를 다양한 언어의 표현으로 가리키고 있다. 즉, 의미 단계(semantic level)에서 정보가 소통되어야, 표층적 표현이 달라도 정보의 호환이나 공유가 가능해진다.



[그림5] 단일 의미의 다양한 표층적 표현

이러한 처리를 위해 W3C는 1999년에 정보 리소스를 의미 단계에서 기술할 수 있는 RDF(Resource Description Framework) 라는 프레임워크를 제시하였다. RDF는 [그림 6-(a)]에서 볼 수 있는 것처럼 모든 사물을 주어(subject), 서술어(predicate), 목적어(object)의 그래프 구조로 표현하고, 개별 주어, 서술어, 목적어에 URI를 부여하여 유일성을 보장하였다.



(a) RDF의 그래프 구조 (b) 사과의 시맨틱 기술

[그림 6] RDF를 이용한 시맨틱 기술

예를 들어, [그림 5]의 사과를 RDF를 이용하면, [그림 6-(b)]와 같은 그래프 구조로 표현이 가능해진다. 이 그래프 구조에서는 사과(ex:apple)은 과일(ex:fruit)의 일종(rdfs:subClassOf)이고 Apple, 사과와 같은 이름을 갖게 된다(ex:hasName). RDF는 그래프 구조이지만, XML, Turtle, JSON-LD와 같은 다양한 데이터 형식으로 표현이 가능하다.

예를 들어, [그림 6-(b)]의 사과 표현의 Turtle 표현은 다음과 같다. 내용을 보면 알 수 있듯이, 그림에서 표현된 그래프의 노드->링크->노드의 관계를 주어, 서술어, 목적어로 표현한 것을 알 수 있다.

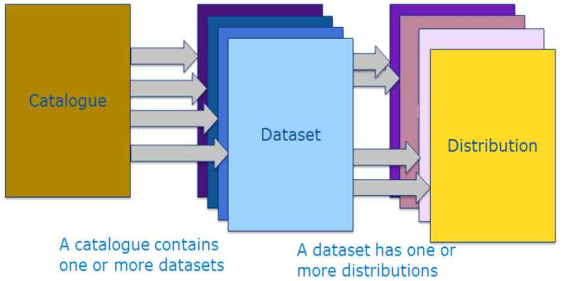
```
@prefix ex: <http://example.com/owl/ex/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

ex:apple rdf:type rdf:Resource
ex:apple rdfs:subClassOf ex:fruit
ex:apple ex:hasName 사과
ex:apple ex:hasName Apple
ex:apple ex:hasName リンゴ
```

(3) DCAT 2.0 분석

DCAT은 W3C가 주도하여 웹에 게시된 데이터 카탈로그 간의 상호 운용성을 용이하게 하기 위해 설계된 RDF 온톨로지(ontology)이며, 2014년에 표준으로 채택되어 있다. DCAT을 사용하여 데이터 카탈로그의 데이터 집합을 설명함으로써 게시자는 검색 가능성을 높이고 여러 카탈로그의 메타 데이터를 쉽게 응용 프로그램에서 사용할 수 있다.

DCAT 1.0에서는 정적인 데이터를 주로 대상으로 고려하여 설계되었기 때문에, [그림 7]과 같이 Catalog가 여러 개의 Dataset을 갖고, 하나의 Dataset이 여러 배포형태인 Distribution을 갖는 구조를 기본으로 갖고 있었다.



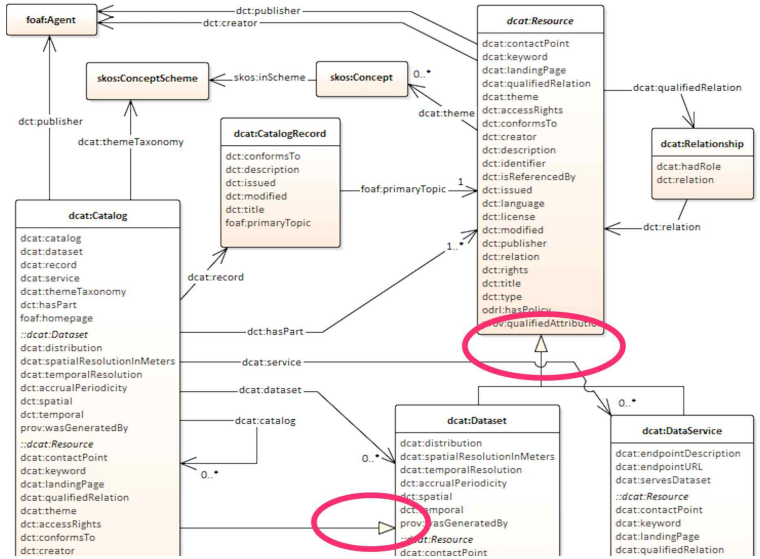
[그림 7] DCAT 1.0에서의 클래스 관계

그러나 DCAT 2.0에서는 동적인 데이터 서비스 지원을 위해, DCAT의 Catalog는 여러 개의 Dataset과 DataService를 가질 수 있도록 수정되었고, Dataset의 배포 형식 표현을 위해 Distribution이 사용된다. 즉, 정적인 파일들을 위해서는 Dataset이 사용되고, 동적 서비스나 API 연동을 위해서는 DataService가 사용된다. [표 1]은 DCAT 2.0의 주요 클래스를 나타내고 있다.

표 1. DCAT 2.0 주요 클래스

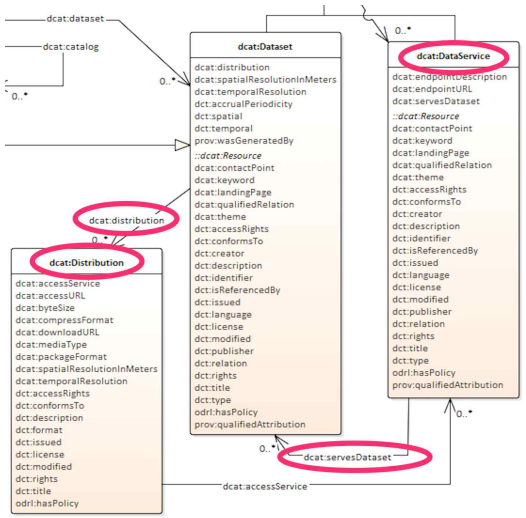
클래스	설명
dcat:Resource	카탈로그에서의 개별 아이템이며, 최상위 클래스
dcat:Catalog	데이터셋이나 데이터서비스를 나타내는 메타데이터들의 집합을 기술하는 카탈로그 클래스
dcat:Dataset	카탈로그에 포함된 데이터 세트를 나타내는 클래스
dcat:Distribution	데이터셋의 배포 형식을 나타내는 클래스
dcat:DataService	API 접근등을 할 수 있는 데이터서비스를 나타내는 클래스

DCAT 2.0에서의 가장 커다란 변화는 Resource 클래스의 도입이다. 이 클래스는 기존의 DataSet과 새로 도입된 DataService의 상위클래스로 공통점을 모아놓은 클래스이다. 또한, Catalog 클래스도 Dataset에서 상속받도록 구성되어 있다. 전반적으로 객체지향 모델링을 조금 더 강조한 형태로 모델링이 되었다. [그림 8]은 이러한 구조를 잘 보여주고 있다.



[그림 8] Resource로부터의 상속

두 번째 변화는 Distribution의 범위를 기존에 비해서 축소하고, 온라인 데이터 전송을 지원하기 위한 DataService를 도입한 것이다. 1.0 버전에서는 Dataset은 데이터에 대한 기본적인 정보를 기술하는 역할을 하고, Distribution은 파일을 제공하는 URL과 전송 파일 형식에 대한 정보를 기술하는 역할에 충실했었다. 이에 비해 DCAT 2.0은 파일 전송과 데이터서비스를 구분하여 구성되어 있다. 파일 전송은 기존의 Distribution의 범위를 축소하여 구성되어 있다. 이에 비해, 온라인 데이터 전송은 DataService 라는 새로운 클래스에서 기술하도록 구성하였다. 이는 앞으로의 데이터 전송은 단순한 파일의 전송보다는 스트리밍 형태의 데이터 전송이 많아질 것으로 예측되기 때문에 이를 반영하여 클래스를 구분한 것으로 볼 수 있다. [그림 9]는 Dataset을 DataService와 Distribution이 나눠서 외부에 제공하는 방식을 기술한 것을 나타낸다.



[그림 9] Distribution과 DataService 의 분리

2021년 5월 현재 DCAT은 3.0 개정 작업의 WD(Working Draft)가 진행되고 있으며, 표준 작업 문서인 WD의 특성상 현재 발표된 내용이 변경될 수 있다. 현재 3.0에서 가장 주요한 움직임은 두 가지로 파악되는데, 첫째는 version을 통한 데이터 계통도를 보완하려는 움직임이며, 둘째는 여러 Dataset의 묶음인 DatasetSeries 클래스의 도입이다. 데이터의 특성상 계속해서 데이터의 추가나 변경이 일어나기 쉽고, 이런 상황에서 판올림 개념이 없으면 기존 데이터와의 관계를 표현하기가 곤란하다. 이러한 한계점을 극복하기 위하여 DCAT 3.0에서는 version과 관련된 version, isVersionOf, hasVersion 등의 속성 추가가 논의되고 있다. 또한 단일 Dataset 외에도 동일 데이터셋이 여러 개 있는 경우를 표현하기 위하여 DatasetSeries 클래스 도입이 제안되고 있으며, DatasetSeries 내에서 Dataset을 구분하기 위한 속성으로 first, last, next, previous 등의 속성이 논의되고 있다.

3. 데이터 품질 표준

(1) 데이터 품질 개요

데이터 품질 관련 작업을 구성하기위한 프레임워크 사고방식을 간략하게 설명하기 위해, DIKW 피라미드라고 알려진 [그림 10]을 참조하면, 데이터 품질은 원시 데이터(data)에서 정보

(information)를 생성하는 과정에서 필요한 기능이라고 정의할 수 있다.

[그림 10] DIKW 피라미드

일반적으로 원시 데이터에는 내포된 형식이나 의미가 없기 때문에 유용하게 사용되려면 해석이 필요하다. 일반적으로 나쁜 데이터 품질은 부정확하고 느린 의사결정으로 이어지며, 좋은 데이터 품질도 시간이 지남에 따라 데이터 품질이 저하되는 경향이 있으며 다른 모든 것과 마찬가지로 열역학 제 2법칙을 따른다. 그리고 서서히 발생한 시스템의 총 엔트로피는 시간이 지나도 절대 줄어들지 않는다. 열역학 제 2법칙은 우주의 어떤 물체도 피할 수 없기 때문에 특별한 노력을 들이지 않으면 데이터 품질 저하는 피할 수 없다. 따라서 문제를 식별하고 데이터 품질을 향상시키는 프로세스를 만들고, 프로세스의 모든 단계에서 데이터 품질 향상 노력을 함으로써 데이터의 가치를 높일 수 있다. 데이터 품질에 대한 많은 정의가 있지만, 일반적으로 데이터의 품질을 정의하는 개괄적인 지표는 다음과 같다.

- **완전성(Completeness)**

데이터가 기대치를 충족하면 완전한 것으로 간주한다. 선택적(임의의) 데이터가 있을 수 있지만 완전성이 부족하다는 것은 원하는 정보가 누락되었음을 의미한다.

- **일관성(Consistency)**

같은 장소에서 동일한 것에 대해 모순되거나 상충되는 정보가 있는 경우 데이터는 일관성이 없는 것으로 간주한다.

- **적시성(Timeliness)**

필요할 때 사용할 수 있는가에 대한 지표이다. 예를 들어, 오류 발생 5시간 후에나 접근 가능하면, 로그 보고서의 데이터로는 적합하지 않다.

- **무결성(Integrity)**

이 개념은 잘 설계된 데이터 및 시스템에 적용된다. 예를 들어 관계형 데이터베이스에서는 고립된 데이터가 없거나 레코드 간의 연결이 깨지지 않음을 의미한다.

- **정확성(Accuracy)**

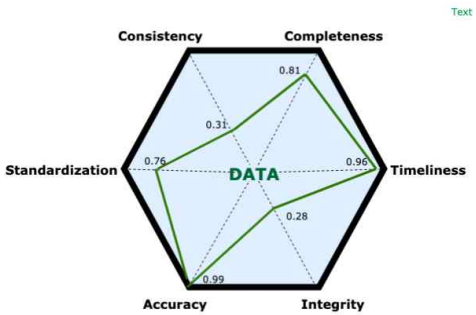
데이터가 실생활에 사용되기에 충분히 정확해야함을 의미한다. 만일 사람의 생일이 13월로 잘못 기록되어 있다면, 이것은 정확성 문제가 있는 것이다.

- **표준화(Standardization)**

이 관점은 여전히 논쟁의 대상이 되지만, 데이터베이스 설계 중에 내린 결정은 일관성이 있어야하며 표준을 따르는 것이 바람직하다. 예를 들어, 관계형 모델에서 다양한 정규화 방식이 표준이다.

데이터 품질의 측정에 관해서는 안타깝지만 아직 일관된 방법이나 정해진 절차는 없으며, 일반적으로 프로젝트에서 요구하는 목적 값을 바탕으로 알고리즘을 만드는 것이 권장된다. 알고리즘은 입력을 받아 각 차원에 점수를 부여 할 수 있다. 예를 들어 [그림 11]에 나타나 바와 같이 각 지표에 대해서 0과 1 사이의 값을 부여할 수 있다. 대량의 데이터를 처리하고 그 점수를 계산하기 위한 다양한 방법이 있고, 또한 데이터베이스의 일부분만을 선택해서 처리할 수도 있다. 그리고 이러한 과정을 자동화하기 위해서는 프로그래밍을 이용하여 결과를 비교하며 동일한 테스트를 반복 실행할 수 있다.

이 중요한 첫 번째 단계를 수행 한 후 각 요소에 가중치를 할당 할 수 있다. 가중치는 데이터를 실제 상황과 더 관련이 있도록 만들기 때문에 중요하다. 일부 데이터 품질 기준은 비즈니스 목표를 달성하는 데 중요 할 수 있지만 다른 기준은 그다지 중요하지 않을 수 있기 때문에 이해관계자와 함께 수행하는 이 분석은 의사결정에 필요한 데이터의 가중치와 우선순위를 정하는데 큰 도움이 된다.



[그림 11] 데이터 품질진단을 위한 지표의 예

(2) W3C 품질 표준

W3C에서는 데이터 품질과 관련하여 Data on the Web Best Practices: Data Quality Vocabulary 표준을 제정하였으며, 이 표준은 널리 활용되고 있다.

● **데이터 품질 어휘 개요**

Data on the Web Best Practices: Data Quality Vocabulary는 2017년에 제정된 표준으로 웹에 게시된 데이터가 제대로 활용되기 위한 여러 모범사례(Best Practices)를 제시하였다. 이 표준에서는 웹에 게시 된 데이터의 품질에 대한 정보 게시의 관련성을 지적하였으며, 이 노력의 일환으로 W3C Data on the Web Best Practices Working Group은 데이터 품질을 표현하기 위한 온톨로지를 생성하였다. 이 표준에 제시된 DQV (Data Quality Vocabulary)는 데이터 품질, 업데이트 빈도, 사용자 수정 수락 여부, 지속성 약속 등을 다루기 위해 DCAT 어휘와 연동 될 수 있다. 웹 콘텐츠 게시자가 이 온톨로지를 사용하면 개발자 사이에서 데이터에 대한 신뢰를 높일 수 있다.

이 온톨로지는 "품질"이 의미하는 바를 결정하지 않는다. 품질은 사용하려는 사람의 관점에서 봐야하기 때문에 객관적이고 이상적인 정의가 없다. 왜냐하면, 일부 데이터셋은 데이터 소비자에게 품질이 낮은 리소스로 판단되지만 동시에 다른 사용자의 요구에는 완벽하게 부합하기 때문이다. 따라서 많은 행위자가 데이터셋의 품질을 평가하고 데이터셋에 대한 주석, 인증서, 의견을 게시할 수 있도록 하는 데 많은 중요성을 부여해야 한다. 데이터셋의 게시자는 데이터 소비자가 데이터셋을 비즈니스에 사용할 수 있는지 여부를 결정하는 데 도움이 되는 메타 데이터를 게시해야한다. 그러나 웹과 같은 개방형 환경에서 게시된 데이터의 품질에 대해 게시자 만이 발언권을 가져서는 안 된다. 인증기관, 데이터 중계인, 데이터 소비자도 관련 품질 평가를 할 수 있어야 한다.

또한 데이터셋 수명 주기의 모든 단계에서 품질 메타 데이터를 보다 쉽게 게시, 교환 및 소비 할 수 있도록 하여 이를 촉진해야 한다. 이것이 품질 측정에서 DQV가 피드백, 주석, 정책 및 인증서에 중점을 두는 이유이다. DQV는 데이터 품질을 나타내는 기존 작업, 특히 연결된 개방형 데이터셋의 품질에 대한 정보 (특히 척도)를 나타내는 daQ 온톨로지에서 영감을 얻고 그와 일치한다.

● **온톨로지의 정의**

W3C에서의 품질 온톨로지는 DCAT을 기반으로 하며, 데이터셋의 품질을 표현하는 데 적합한 여러 추가 속성 및 클래스로 확장된다. 주어진 데이터셋 또는 분포의 품질은 다수의 관찰된 속성을 통해 평가된다. 예를 들어, 특정 표준을 준수하기 때문에 데이터셋이 고품질이라고 간주 할 수 있지만, 다른 사용 사례의 경우 데이터 품질은 다른 데이터셋과의 상호 연결 수준에 따라 달라진다. 이러한 속성을 표현하기 위해 dcat:Dataset 또는 dcat:Distribution의 인스턴스는 다음 [표 2]의 클래스로 표현되는 5가지 다른 유형의 품질 정보와 관련 될 수 있다.

[표 2] 5가지 유형의 품질 정보

어휘	내용
dqv:QualityAnnotation	데이터셋 또는 해당 배포에 대해 제공된 피드백 및 품질 인증서를 나타낸다.
dcterms:Standard	데이터셋 또는 해당 배포가 준수하는 표준을 나타낸다
dqv:QualityPolicy	주로 데이터 품질 문제에 의해 관리되는 정책 또는 계약을 나타낸다.
dqv:QualityMeasurement	데이터셋 또는 분포에 대한 정량적 또는 정성적 정보를 제공하는 척도 값을 나타낸다.
prov:Entity	데이터셋 또는 배포의 출처와 관련된 엔티티를 나타낸다.

이 외에도 소비자와 관련된 데이터셋의 품질 관련 특성을 기술하기 위해, 차원(dqv:Dimension)이 정의되었고, 구체적인 품질 지표를 관찰하여 추상적인 데이터 품질 차원을 측정하는 절차를 제공하기 위해 척도(dqv:Metric)가 정의되었다. 통상적으로 여러 개의 척도가 하나의 차원에 속하게 된다. 예를 가용성(availability) 차원은 SPARQL 엔드 포인트, 또는 RDF 덤프의 접근성의 척도로 나타낼 수 있다. 이때의 척도 값은 숫자 또는 boolean 형이 될 수 있다. 즉, 차원이 개념적인 의미라면 차원에 속한 척도는 측정 가능한 구체적인 값이라고 할 수 있다. 품질 측정 외에도 DQV는 차원에 따라 구성 할 수 있는 인

증서, 표준 및 품질 정책을 고려한다. [그림 12]는 DQV에서 새로 정의된 어휘와 기존 DCAT 어휘와의 관계를 보여준다.

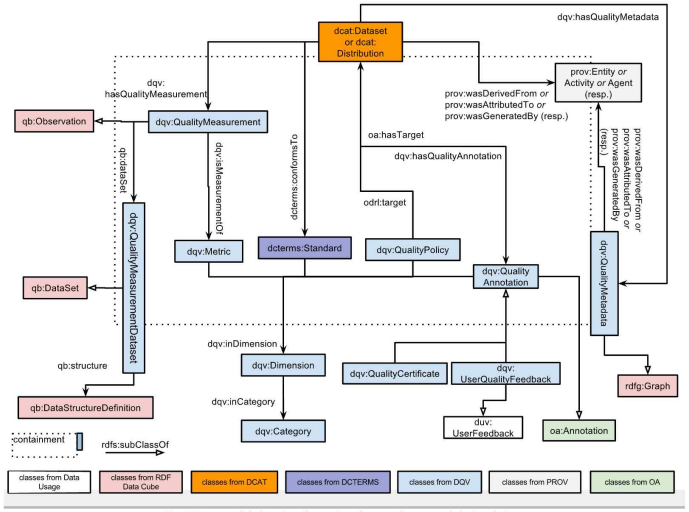


Fig. 1 Data model showing the main relevant classes and their relations.

[그림 12] DQV 어휘와 기존 DCAT 어휘의 관계

● **품질 표준의 사용 예**

데이터 품질을 처리하기 위한 DQV는 EU 데이터 포털에서 실제로 활용이 되고 있다. 보유한 모든 데이터셋에서 [그림 13]과 같이 품질과 관련된 DQV Data를 얻을 수 있으며, 이를 통해 데이터셋의 품질 관리를 하고 있다.

[그림 13] EU 데이터 포털에서의 DQV 제공

제공되는 DQV 데이터는 XML, Turtle, JSON-LD와 같은 다양한 데이터 포맷으로 데이터를 제공하고 있으며, [그림 14]처럼 DQV 척도를 기술하고 있다.

```
<http://www.w3.org/ns/dqv#hasQualityMeasurement>
  [ rdf:type <http://www.w3.org/ns/dqv#QualityMeasurement> ;
  <http://www.w3.org/ns/dqv#computedOn>
    <https://europeandataportal.eu/set/distribution/4f7572c4-05dd-4b8c-82e7-ed6e7b15702b> ;
  <http://www.w3.org/ns/dqv#isMeasurementOf>
    <https://piveau.eu/ns/voc#downloadUrlAvailability> ;
  <http://www.w3.org/ns/dqv#value>
     true ;
  <http://www.w3.org/ns/prov#generatedAtTime>
     "2021-05-21T16:57:52.867655Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>
   ];
```

[그림 14] 품질 데이터 DQV의 Turtle 표현식 일부

4. 결론 및 시사점

인공지능이 전 산업 영역으로 빠르게 확산되고 이를 통한 사회변화가 가속화되는 상황에서 인공지능 성능의 토대가 되는 양질의 데이터 확보는 4차 산업혁명 산업 발전을 위해서 무엇보다 시급한 문제가 되었다. 이러한 상황에서 데이터 유통을 촉진하는 메타데이터 표준인 DCAT 표준의 이해와 적용은 매우 중요하다고 할 것이다. 또한 단순하게 많은 양의 데이터를 빠르게 확보하는 것이 관건이었던 빅데이터 분야와 달리 인공지능은 학습에 사용되는 데이터의 품질에 따라, 예측 성능에 큰 차이가 있기 때문에 데이터의 품질을 높이기 위한 노력이 필수적이라고 하겠다.

이러한 요구사항 때문에 각국은 데이터 생성과 유통을 가속화하고자 하는 노력을 하고 있으며, EU 데이터 포털의 사례에서도 볼 수 있듯이 데이터 유통과 데이터 품질 표준을 적극 활용하고 있다. 국내에서도 한국지능정보사회진흥원이 추진하는 통합데이터 지도(https://www.bigdata-map.kr) 사이트가 DCAT 2.0을 기반으로 구축된 것은 매우 바람직한 방향이라 생각된다. 또한 그래프 검색을 제공하여 데이터 간의 관계를 보여준 점은 데이터 활용도를 높이는 좋은 방안이라 생각된다. 다만, 추후에 EU 데이터 포털과 같이 데이터 품질과 관련된 내용을 DQV 표준에 맞춰서 보완한다면 데이터 품질 향상에 도움이 될 것으로 판단된다. 그리고 DCAT 3.0의 경우에 판올림 기능이 추가되었는데, 데이터 포털의 지속 가능성에 도움이 될 수 있으므로 지원 여부가 고려되어야 할 것이다.

참 고 문 헌

[1] W3C, Recommendation 16, “Data Catalog Vocabulary (DCAT)”, 2014. 1.

[2] W3C, Recommendation 4, “Data Catalog Vocabulary (DCAT) - Version 2”, https://www.w3.org/TR/vocab-dcat-2/ 2020. 2.

[3] W3C, Recommendation 16, “Data Catalog Vocabulary (DCAT) - Version 3”, https://www.w3.org/TR/vocab-dcat-3/ 2021. 5.

[4] Bert Van Nuffelen, “DCAT Application Profile for data portals in Europe Version 2.0.1”, 2020. 6.

[5] W3C, Recommendation 31, “Data on the Web Best Practices”, https://www.w3.org/TR/2017/REC-dwbp-20170131/ 2017. 1.