빅데이터 유통을 위한 카탈로그 메타데이터 표준 동향 및 서비스 구조

정의현 _ 안양대학교 융합소프트웨어학과 교수

1. 머리말

빅데이터 환경에서 데이터는 새로운 가치를 창출하는 새로운 '자원'으로 주목받고 있으며, 4차 산업혁명 경제에서 핵심적인 역할을 할 것으로 기대되고 있다. 전 세계 빅데이터 시장 매출은 2018년 420억 달러에서 2027년에는 1,030억 달러로 증가할 것으로 예측되고 있으며, 대기업 임원의 79%가 빅데이터를 수용하지 않는 기업은 경쟁력을 잃고 멸종 위기에 처할 수 있다는 데 동의한다고 미국의 다국적 경영 컨설팅 기업인 액센추어(Accenture)는 보고하고 있다. 그러나 이렇게 밝은 전망과는 달리 아직도 많은 기관과 기업에서 생성되는 빅데이터는 적절한 시점에 분석되지 못하고, 미래의 분석을 위해 단순히 저장만 하는 단계에 머무르고 있는 것이 현실이다. 일례로 IDC 보고서에 의하면 세계적으로 현재 0.5% 데이터만이 분석되고 있는 것으로 추정되고 있다.

데이터가 흘러넘치도록 많은데도 불구하고 제대로 활용되지 못하는 가장 큰 이유는 아이러니하게도 데이터 분석을 위한 양질의 데이터 확보가 어렵다는 점이다. 많은 경우 데이터의 생산자는 소비자나 분석가가 아니기 때문에 데이터는 생산 후 방치되게 되고, 소비자나 분석가는 필요한 데이터를 구할 수 없어서 데이터 비즈니스를 수행할 수 없는 모순적인 상황에 처하게 된다. 따라서 저장매체에 방치되어 있는 빅데이터를 제대로 활용하기 위해서는 데이터의 생산자와 소비자를 이어주는 중간 거래(broker) 기능, 즉 데이터의 이동에 관계되는 모든 경제 활동 구성원들이 '서로 협력하는 가치사슬 연합군'인 빅데이터 유통 생태계가 매우 중요한 요소로 대두되게 된다.

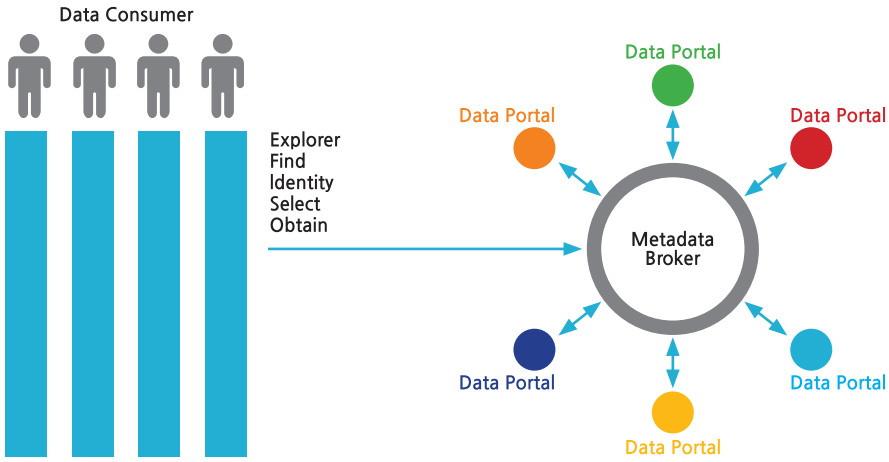
주요 선진국들은 빅데이터 유통 생태계가 빅데이터 활성화의 핵심이라는 것에 주목하고 데이터를 효율적으로 유통하기 위한 다양한 방안을 제시하고 있다. 영국의 '오픈 데이터 정책'과 같은 것이 이러한 노력의 일환이며, 특히 EU의 경우 다양한 국가가 하나의 연합체에 들어있기 때문에 개별 국가에서 산출된 데이터 공유나 유통에 대한 노력이 수행돼 왔다. 이러한 노력은 'European Data Portal'이라는 데이터 포털을 기반으로 한 생태계 구축으로 이어졌고, 이 데이터 포털을 통해 EU의 공공 부문 데이터가 유통되고 있다.

이렇듯 데이터의 공유를 통한 빅데이터 활성화를

위해서는 데이터 포털이 매우 중요하나, 아쉽게도 데이터 포털의 구축만으로 빅데이터 활성화가 되지 않는다는 점이 최근에 지적되고 있다. 기본적으로 빅데이터는 일반 데이터와 달리 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 빅데이터 유통 생태계에서는 개별 빅데이터의 특성을 제시하는 메타데이터(metadata)의 도입이 필연적이다. 일반적으로 빅데이터를 위한 메타데이터는 기존의 다양한 메타데이터 유형 및 비정형 데이터를 위한 최소한의 코어 메타데이터를 제공함으로써, 이들을 통합·연계·관리할 수 있어야 한다.

따라서 이러한 메타데이터의 도입은 빅데이터의 유통이나 생태계 활성화 측면에서 권장되어야 함에도 불구하고, 데이터 포털별로 개별 메타데이터를 정의하는 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 된다. 2014년 조사에 의하면, EU에서만 160개의 데이터 포털이 존재하고 있다고 알려져 있으며, 국내에서도 주요 공공기관, 지자체, 이익단체별로 데이터 포털이 존재하고 있는 상황이다. 하지만, 거래소마다 각자의 비즈니스 전략과 데이터 관리 방식이 상이하기 때문에 모든 데이터 포털에 특정한 메타데이터 형식을 강제하는 것도 현실적으로 어려운 문제이다.

이러한 문제를 해결하기 위하여 W3C에서는 2014년에 빅데이터 유통에 필요한 새로운 메타데이터 표준인 DCAT(Data Catalog Vocabulary)을 제시하였으며, 2015년에는 유럽연합 집행위원회(EU Commission) 주도로 DCAT-AP(DCAT Application Profile) 표준이 제정되었다. 두 표준은 현재 EU에서 채택되어 European Data Portal의 기본 메타데이터로 사용되고 있으며, [그림 1]과 같이 여러 데이터 포털의 호환을 제공하는 서비스 모델을 제시하고 있다. 본고에서는 DCAT을 중심으로 빅데이터 유통을 위한 카탈로그 메타데이터의 표준 동향 및 실제 적용 사례에 대해 살펴보도록 한다.



[그림 1] European Data Portal에서의 메타데이터 호환

2. DCAT

2.1. DCAT의 특징

DCAT은 W3C가 주도하여 웹에 게시된 데이터 카탈로그 간의 상호 운용성을 용이하게 하기 위해 설계된 RDF(Resource Description Framework) 온톨로지(ontology)이다. 개별 포털에서 기존 메타데이터를 변경할 필요 없이 DCAT으로 메타데이터 카탈로그를 만들어 배포하면 검색 노출 가능성이 높아지고, 기존 데이터 제공자들이 이미 투자한 정보 자산을 포기하지 않고도 데이터 유통 생태계로 편입되는 효과를 낳게 한다. 즉, 개별 데이터 포털에서 생성되는 데이터의 메타데이터는 존중하면서, DCAT이나 DCAT-AP로 메타데이터 카탈로그를 만들면 해당 데이터들의 호환을 제공한다는 점이다.

DCAT의 가장 큰 특징은 시맨틱웹 기술을 이용한 RDF 온톨로지로 구성되어 있다는 점과 개별 요소들이 카탈로그(Catalog)를 중심으로 계층적 구조를 갖고 있다는 점이다. 먼저, DCAT은 RDF 온톨로지로 구성되어 있기 때문에 시맨틱웹의 장점인 의미(symantic) 수준의 데이터 호환이 가능하므로 개별 데이터 포털에서 각자 메타데이터를 정의해도 의미 충돌이 나지 않는 장점을 갖게 된다. 또한 시맨틱 검색이 가능해져서 보다 정확한 데이터 검색이 가능해진다.

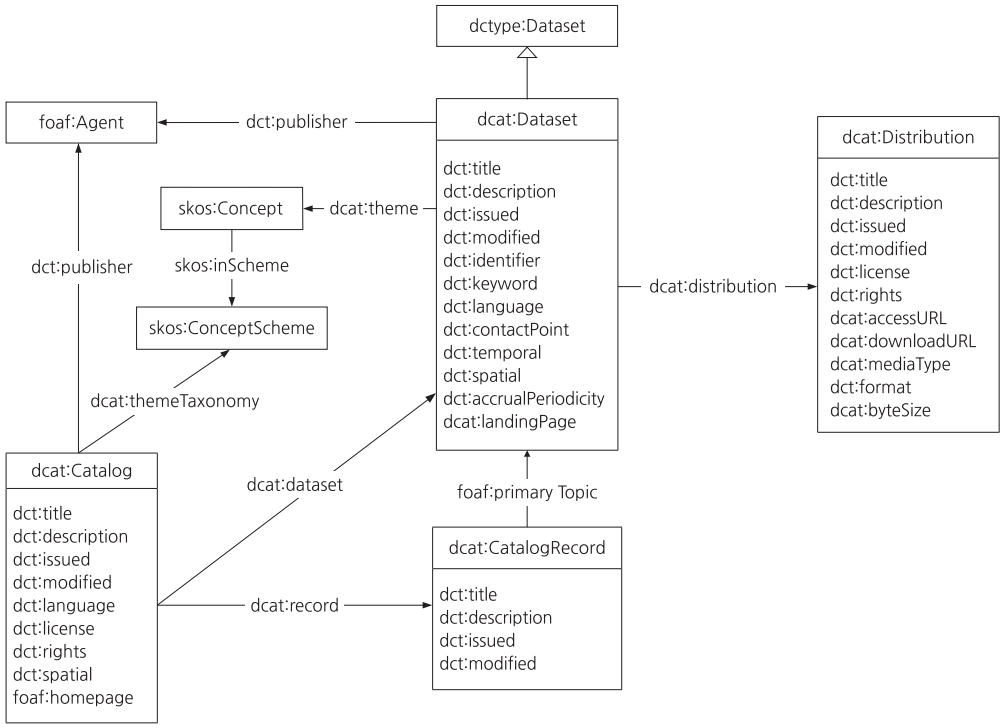
두 번째 특징으로, DCAT은 하나의 분야에 해당하는 Catalog 클래스가 카탈로그에 포함된 여러 개의 데이터세트(Dataset)에 대응되는 Dataset들로 구성되고, 개별 Dataset은 배포와 관련한 여러 개의 Distribution을 가지는 계층적 구조를 갖는다는 것이다. 예를 들어, 의료보험 관련 데이터를 DCAT을 이용해서 정의한다면, 의료보험에 대한 카탈로그는 HealthInsurance라는 Catalog 클래스를 이용하여 정의할 수 있다. 그리고 처방과 약과 관련된 데이터세트는 각각 Prescription, Medicine 등의 Dataset 클래스로 정의될 것이다. 그리고 Medicine 데이터세트가 JSON 다운로드와 CSV 배포를 지원한다면, 해당하는 2개의 Distribution 클래스를 정의해서 기술될 수 있다. [그림 2]는 DCAT 온톨로지의 UML 다이어그램 형태를 보여준다.

2.2. DCAT 개정판

W3C에 의해 DCAT은 2019년 현재 개정판이 준비되고 있으며, 기존 DCAT에 비해서 몇 가지 변화가 있을 것으로 예측된다.

첫째, Resource라는 최상위 클래스를 정의하여 Catalog, Dataset, Distribution이 공통으로 가지고 있었던 속성들을 갖고 있도록 하고, Catalog, Dataset, Distribution은 모두 Resource로부터 상속받아서 사용하도록 정의되었다. 이러한 변화로 인하여 기존 DCAT이 갖고 있었던 계층적 구조의 일관성은 다소 약화되었으나 개별 클래스를 Resource라는 공통의 클래스를 통해서 접근할 수 있는 장점이 생겼고 메타데이터를 보다 객체지향적으로 만들 수 있게 되었다.

둘째, 배포방식의 정의가 다소 모호했던 Distribution을 파일 배포용으로 범위를 축소해서 정의하고, 대신에 API(Application Programming Interface)나 SaaS(Service as a Service)와 같은 온라인 접근을 정의할 수 있는 DataService가 새롭게 정의되었다. 이는 최근 데이터 제공 방법이 파일을 제공하는 것 외에 API나 SaaS와 같이 서비스 접근을 통한 제공이 확산되면서 이를 지원하기 위한 것으로 판단된다. DCAT 개정판은 이러한 변화를 통해 기존 DCAT보다 세밀한 메타데이터 모델링이 가능할 것으로 기대된다.

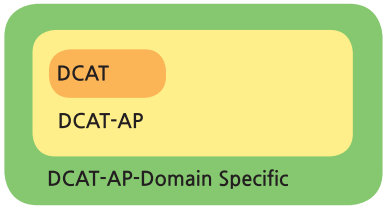


[그림 2] DCAT의 UML 다이어그램

3. DCAT-AP

DCAT 기반의 프레임워크 설계는 DCAT 대부분의 장점을 이용할 수 있다는 특징이 있으나, 데이터 카탈로그의 일반적인 요소가 강조되어 있기 때문에 특정 응용 분야(특히 데이터 교환)에 적합하도록 구성된 프로파일(Profile)이 요구되는 경우에는 조금 더 세밀한 설계가 필요하게 된다. 이러한 목적을 위해서 EU를 중심으로 DCAT의 Application Profile을 정의하는 연구가 진행되었고, DCAT-AP라는 표준안이 2015년에 제시되었다. 이 표준안은 공공 부문의 데이터 교환을 주목적으로 하였고, EU 역내 국가 간의 공공 데이터 교환을 위한 기본 구조로 활용되었다.

DCAT-AP는 DCAT을 기반으로 데이터 교환을 목적으로 클래스와 프로퍼티를 추가하고, 각 클래스/프로퍼티의 필수(Mandatory)/권고(Recommended)/선택(Optional)을 지정하여 보다 명확한 데이터 검색과 유통이 가능하도록 하였다. 아래 그림은 DCAT, DCAT-AP, 분야별 DCAT-AP의 관계를 보여준다.



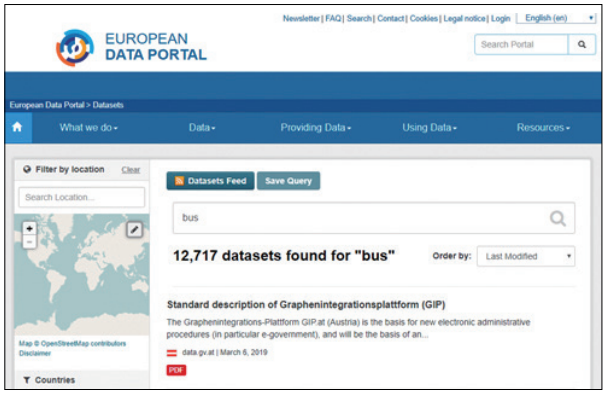
DCAT-AP의 경우에는 분야별 AP의 정의를 지원하기 때문에 통계, 지리정보, 대중교통 등에서 이미 개별 AP가 정의되어 있다.

- **StatDCAT-AP:** 공개 통계 데이터에 대해 공통적으로 합의된 보급 어휘를 제공하는 것을 목표로 한다. StatDCAT-AP는 DCAT-AP 모델에 일정한 수의 추가 기능을 정의하여 통계 데이터 교환 표준인 SDMX(Statistical Data and Metadata eXchange)와 같은 형식의 데이터 집합을 기술하는 데 사용할 수 있도록 설계하였다.
- **GeoDCAT-AP:** DCAP-AP의 확장으로 지리정보 데이터세트와 데이터세트 시계열, 서비스를 기술한다. GeoDCAT-AP는 ISO 19115와 ISO 19119에 기반한 INSPIRE 메타데이터 규정[INSPIRE-MD-REG]이나 INSPIRE 메타데이터 기술 가이드라인[INSPIRE-MD-TG]을 대치하지 않는다. 이 AP의 주요 목적은 지리정보 메타데이터 소유자들이 RDF를 이용해서 호환성을 높이게 하는 것이다. 2015년에 표준으로 제정되었다.
- **기타 AP:** 이외에도 다양한 AP가 만들어지고 있는데, 대중교통에 대한 정보를 표현하는 TransportDCAT-AP와 연구 데이터 유통을 위한 JRC(Joint Research Centre) 메타데이터가 AP로 연구되고 있다.

4. EU에서의 DCAT/DCAT-AP 적용 사례

EU의 경우 여러 국가와 기관에서 데이터 포털 서비스를 제공하기 때문에 이의 호환이 중요한 문제로 대두되었다. EU는 이를 해결하기 위해 [그림 3]과 같은 European Data Portal을 제시하였다.

European Data Portal은 EU의 모든 데이터를 하나의 포털에서 보유하는 전략을 택하지 않았다. 개별 데이터는 각 기관의 데이터 포털이 유지하고 대신에 데이터에 대한 카탈로그만을 European Data Portal에 유지하는 방법을 채택한 것이다. 예를 들어, European Data Portal도 일반적인 데이터 포털과 마찬가지로 카테고리별로 데이터를 제공하고, 검색은 SPAQL과 키워드 검색을 제공한다. 따라서 표면상으로는 다른 데이터 포털과 차이가 없는 것처럼 보이지만, 검색어를 통해서 데이터를 찾으면, 그 결과는 기존 데이터 포털과 다른 결과물을 보여준다. 기존의 데이터 포털들은 자신이 보유하고 있는 데이터의 내용과 사용 방법을 알려주는 데 반해, European Data Portal은 실제 데이터를 보유하고 있는 데이터 포털에서 받은 카탈로그 정보를 보여주고, 해당 데이터 포털의 데이터로 연결되는 링크를 제시한다. [그림 3]에서 bus라는 키워드로 검색하면, bus 관련 데이터를 갖고 있는 유럽 여러 나라의 데이터 포털을 보여주는 것을 볼 수 있다. 즉, European Data Portal은 여러 데이터 포털들의 카탈로그 데이터 포털인 셈이다.



[그림 3] European Data Portal

5. 맺음말

빅데이터가 성공하기 위해서는 데이터 전문가의 양성, 데이터 생태계 구성, 규제 철폐 등의 여러 요소들이 필요하지만 무엇보다 선결조건은 양질의 데이터 확보 문제이다. 데이터 분석 분야의 오랜 격언인 'Garbage In, Garbage Out'은 빅데이터 분야에서도 여전히 유효한 격언이며 양질의 데이터 확보 문제는 빅데이터 성공의 시금석이 되고 있다. 그러나 데이터 공유에 대한 인식 부족과 데이터 관리 실패로 인한 오염된 데이터 생성 등의 다양한 문제로 인하여 연구자들에게 데이터 확보는 큰 난관이 되고 있다. 또한, 데이터 과학자의 수급 부족으로 인하여 우수한 데이터 과학자를 기업들과 연구기관들이 확보하기 쉽지 않은 것도 빅데이터 확산에 걸림돌이 되고 있다.

이를 해결하기 위해서는 각 기관에서 활용 가능한 데이터를 공개하여 개별 기업들에게는 데이터 확보 문제를 해결하고, 데이터를 분석할 수 있는 데이터 과학자들이 쉽게 데이터에 접근할 수 있게 하는 것이 중요하다. 그러나 단순히 기관별로 데이터 포털을 많이 만드는 것은 데이터 파편화 현상을 가속하기 때문에 바람직한 접근 방안이 아니다. 이보다는 데이터 생산자와 소비자 간의 데이터 유통을 원활히 지원하는 빅데이터 유통 시스템, 즉 빅데이터 유통 생태계의 구축에 대한 전반적인 고민이 필요한 시점이다.

EU의 사례를 보면 기존 데이터 포털들의 자율성을 보장하면서도 데이터 검색과 유통을 원활하게 하는 DCAT, DCAT-AP와 같은 데이터 카탈로그 메타데이터 표준을 만들어 빅데이터 유통 생태계를 효과적으로 구축하였다. 이는 빅데이터 유통 생태계를 구축하기 위해 어떻게 접근해야 하는지에 대한 좋은 참고사례가 될 수 있으며, 여러 측면에서 시사하는 바가 크다 할 것이다. 향후 국내 빅데이터 산업의 활성화를 위해서는 국제 표준과 호환되면서도 국내 상황을 고려한 카탈로그 메타데이터 표준에 대한 연구가 시급히 진행되어야 할 것이며, 이를 적용한 카탈로그 메타데이터 포털의 구성에 대한 정책적 노력이 필요할 것이라고 제언한다.

[참고문헌]

[1] '세계속의 빅데이터', Bigdata Monthly: 빅데이터 동향과 이슈, 한국정보화진흥원, Vol. 42, 2018. 6.

[2] Introduction to metadata management, Open Data Support, 2014.

[3] Data Catalog Vocabulary (DCAT), W3C, 2014.

[4] DCAT Application Profile for data portals in Europe. ISA Programme of the European Commission, 2015.