JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 1, pp. 59-65, Jan. 2018

오픈데이터 플랫폼의 상호운용성을 위한 DCAT 기반 메타데이터 변환도구 설계 및 구현

박경현[1] · 원희선[1] · 류근호[2*]
[1]한국전자통신연구원 스마트데이터연구그룹
[2]충북대학교 데이터베이스/바이오인포매틱스 연구실

A Design and Implementation of a DCAT-based Metadata Transformation Tool for Interoperability in Open Data Platforms

Kyounghyun Park[1] · Hee Sun Wonk[1] · Keun Ho Ryu[2*]
[1]Smart Data Research Group, ETRI, Daejeon, 34129, Korea
[2*]Database/Bioinformatics Lab., Chungbuk National University, Cheongju, 28644, Korea

[요 약]

공공데이터가 국가 경제발전의 자원으로 인식되기 시작함에 따라 세계 각국에서는 공공데이터 포털을 구축하여 민간에게 공공데이터를 개방하기 시작하였다. 이러한 흐름에 맞추어 오픈소스 진영에서도 CKAN을 선두로 오픈데이터 플랫폼 기술이 발전하기 시작하였고 메타데이터의 표준 기술을 적용함으로써 타 플랫폼과의 메타데이터 연동도 가능해지게 되었다. 하지만 아직도 많은 세계 각국의 정부와 지방 자치단체들이 공공데이터 포털을 자체적으로 개발하여 서비스를 하고 있는 실정이기 때문에 각 공공데이터 포털들간의 데이터 공유가 어려운 실정이다. 이에 본 논문에서는 이러한 문제점을 해결하기 위하여 DCAT 기반의 메타데이터 변환도구를 설계, 구현하고 데이터셋을 메타데이터 표준인 DCAT으로 변환하는 방법을 소개한다.

[Abstract]

As open data(public data) began to be recognized as a source of national economic development, many countries began to build public data portals and provide open data to the private sector. In accordance with this trend, open source communities have begun to develop open data platform such as CKAN and enable to share dataset among open data platforms by applying metadata standard technology. However, many governments and local governments are still making it difficult to share data between data portals because they build their own platforms. In this paper, we propose a DCAT-based metadata transformation tool to solve these problems, and show how to transform a dataset into DCAT.

색인어 : 공공데이터, 오픈데이터 플랫폼, CKAN, DCAT, 데이터 유통

Key word : Public Data, Open Data Platform, CKAN, DCAT, Data Delivery

http://dx.doi.org/10.9728/dcs.2018.19.1.59

Received 31 December 2017 ; **Revised** 23 January 2018
Accepted 29 January 2018

***Corresponding Author; Keun Ho Ryu**

Tel: +82-43-267-2254
E-mail: khryu@chungbuk.ac.kr

Ⅰ. 서 론

딥러닝(deep learning)과 인공지능(AI) 기술의 발달로 데이터 분석에 대한 관심이 증가하면서 사회 전분야에 걸친 공공데이터의 중요성이 증가하게 되었고 일부 국가에서는 정부기관 및 지방자치단체를 중심으로 민간 기업들이 공공데이터를 활용하여 이윤을 창출할 수 있도록 오픈 데이터 정책을 수립하고 시행하기 시작하였다.

공공데이터는 일반적으로 공공기관에 의해 발생한 데이터로 기상 정보로부터 신용평가 정보에 이르기까지 사회 전분야에 걸친 오픈데이터를 의미한다. 공공데이터의 중요성은 오픈데이터 플랫폼이 확산, 발전함에 따라 더욱 높아졌는데 대표적인 오픈데이터 플랫폼으로 오픈소스 진영에서는 CKAN (Comprehensive Knowledge Archive Network)[1]과 DKAN[2]이 있고 상용 플랫폼으로는 Socrata[3]가 있다.

오픈데이터 플랫폼은 데이터셋의 메타데이터 정보를 통해 데이터셋을 관리하고 다양한 검색기능을 제공한다. 또한 오픈데이터 플랫폼간의 정보교환도 메타데이터를 통해 이루어지기 때문에 오픈데이터 플랫폼의 활용 및 데이터 공유에 있어서 메타데이터의 표준화 지원은 오픈데이터 플랫폼의 활성화에 있어서 가장 중요한 요소중의 하나이다.

공공데이터 유통과 관련된 메타데이터 표준 기술중 가장 대표적인 기술로는 웹상의 데이터를 통합 관리하기 위해 제정된 데이터 카탈로그 표준인 DCAT(Data Catalog Vocabulary)[4]이 있다. DCAT은 W3C에서 제정한 데이터 카탈로그 표준으로 웹상의 다양한 데이터 소스를 기술하는 메타데이터를 RDF 형태로 정의함으로써 데이터의 접근 및 활용을 가능하게 해주는 표준기술이다. DCAT은 CKAN, DKAN, Socrata와 같은 대표적인 오픈데이터 플랫폼에서 지원하고 있으며 실제로 CKAN을 기반으로 하는 대표적인 공공데이터 포털인 data.gov[5]와 data.gov.uk[6]에서도 데이터 공유를 위해 DCAT 하베스팅 기능을 제공하고 있다.

DCAT의 데이터 모델은 7개의 클래스로 구성된다. 특히 3개의 주클래스인 Catalog, DataSet, Distribution 클래스를 포함하고 1개의 중요클래스인 CatalogRecord 클래스를 포함한다.

대표적인 오픈데이터 플랫폼인 CKAN은 CKAN 하베스팅을 통해 플랫폼간에 데이터셋 정보를 공유한다. 또한 DCAT 하베스팅을 지원함으로써 Socrata와 같이 이기종의 플랫폼과도 데이터셋 정보를 공유할 수 있다.

하지만 많은 정부기관과 지방자치단체들이 오픈데이터 플랫폼을 그대로 활용하기도 하지만 요구사항에 의해 자체 개발하거나 기존 플랫폼들을 수정, 확장하는 경우도 많이 존재한다. 그 예로, 한국에서도 대표적인 공공데이터 포털들이 대부분 자체 개발되어 공공데이터를 서비스하고 있다[7],[8]. 이와 같은 경우 데이터 포털간에 데이터셋 정보를 상호교환할 수 없다는 문제가 발생한다.

이에 본 논문에서는 이러한 문제점을 해결하기 위해 데이터 포털간 상호운용성을 지원해주기 위한 방법으로 DCAT 기반의 메타데이터 변환도구를 소개한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 주요 오픈데이터 플랫폼을 소개하고 W3C에서 제정한 메타데이터 표준인 DCAT에 대해 설명한다. 3장에서는 메타데이터 변환도구에서 DCAT 기반의 메타데이터 변환을 위해 데이터셋을 관리하는 데이터스토어(data store)에 접근하는 방법과 데이터셋과 DCAT과의 매핑하는 방법에 대해 기술하고 메타데이터 변환도구를 통해 변환된 RDF 파일을 저장, 배포하는 과정에 대해 기술한다. 마지막으로 4장에서는 결론으로 메타데이터 변환도구의 적용사례와 향후 연구방향을 기술한다.

Ⅱ. 관련 연구

2-1 오픈데이터 플랫폼

오픈데이터 플랫폼은 공공데이터의 유통 관점에서 볼 때 데이터 소유자에게 데이터를 등록하고 공개, 유통할 수 있는 기능을 제공하고 데이터 사용자에게는 데이터를 효율적으로 검색하고 활용할 수 있게 하기 위한 기능을 제공하는 데이터 유통 플랫폼으로 정의할 수 있다.

따라서 오픈데이터 플랫폼이 데이터 유통 플랫폼으로의 기능을 수행하기 위해서는 다음과 같은 기능을 제공해야 한다.

첫째, 데이터 등록 기능을 제공해야 한다. 즉 데이터 소유자가 데이터를 업로드하는 기능을 제공해야 한다. 데이터를 업로드할 때 데이터 소유자는 데이터셋과 함께 데이터셋에 대한 세부정보들(메타데이터)을 업로드할 수 있다.

둘째, 데이터 발행 기능을 제공해야 한다. 데이터 발행을 통해 데이터 사용자들은 활용 가능한 데이터셋들을 검색할 수 있다.

셋째, 오픈데이터 플랫폼은 데이터 현황관리 기능을 제공해야 한다. 데이터 현황관리란 데이터셋의 세부정보를 관리하는 기능으로 메타데이터 관리 또는 데이터 카탈로그 관리라고도 한다. 데이터의 주요 세부정보로는 데이터의 출처, 분류, 내용 및 데이터의 이력정보등이 있다.

넷째, 오픈데이터 플랫폼은 데이터 포털 기능을 제공해야 하는데 사용자는 포털을 통해 데이터 유통에 필요한 전반적인 기능들을 사용할 수 있는 환경을 제공받을 수 있다.

마지막으로, 시각화 기능을 제공해야 한다. 시각화 기능이란 데이터 및 데이터 분석 결과를 도표, 그래프등을 이용하여 시각적으로 제공하는 기능으로 특히 오픈소스와 비교할 때 상용 플랫폼에서 풍부한 시각화 기능을 제공한다.

이와 같은 기능을 포함하는 오픈데이터 플랫폼은 전세계적으로 공공데이터 포털을 구축하는데 활용되어 왔고 그중에서도 공공데이터 포털을 구축하기 위한 대표적인 오픈데이터 플랫폼으로는 CKAN, DKAN, OGPL(Open Government Platform)[9], Socrata등이 있다.

CKAN은 이중 가장 많이 활용되는 오픈소스 기반의 오픈데이터 플랫폼으로 비영리단체인 OKF(Open Knowledge Foundation)의 주도하에 개발이 진행되고 있다. 현재는 영국, 미국, 캐나다등 40여개국에서 사용하고 있으며 기본적으로 데이터 검색 기능, 키워드 유사 매칭 기능, CSV 데이터 시각화 기능 등을 제공하고 기능을 확장하기 위한 방법으로 다양한 플러그인(extension)을 추가할 수 있도록 설계되었다. 예를 들면, CKAN에서 하베스팅 기능은 기본 기능으로 제공되지 않고 CKAN 하베스팅과 DCAT 하베스팅 플러그인을 추가로 설치함으로써 하베스팅 기능을 제공할 수 있다.

DKAN은 Drupal을 기반으로 만들어진 오픈데이터 플랫폼으로 미국의 NAUM의 주도하에 개발된 플랫폼이다. DKAN은 CKAN의 기능 대부분을 포함하고 있고 프로그램 언어와 데이터베이스지원 같은 기술적인 측면에서 차이점을 보이고 있다.

OGPL은 미국과 인도 정부에서 공동으로 개발한 오픈데이터 플랫폼으로 미국과 인도 정부의 오픈데이터 플랫폼의 장점만을 모아 개발하고 있는 오픈데이터 플랫폼이다.

Socrata는 클라우드 기반의 오픈데이터 플랫폼으로 미국의 여러주에서 사용중인 데이터 플랫폼이다. Socrata의 특징으로 SODA API와 SOQL(SODA Query Language)를 지원하고 성능 향상을 위해 Truth 스토어를 지원한다. 이외에도 Socrata는 오픈소스 기반의 데이터 플랫폼에 비해 풍부한 분석 및 시각화 기능을 제공하고 있다. 하지만 오픈소스인 CKAN이 전세계적으로 사용이 확대됨에 따라 Socrata도 활성화를 위해 오픈소스로 개방하려는 노력을 진행중이다.

2-2 DCAT(Data Catalog Vocabulary)

오픈데이터 플랫폼은 데이터 카탈로그를 활용하여 데이터셋을 관리하고 다양한 검색기능을 제공한다. 또한 다른 플랫폼과의 정보교환 및 검색도 데이터 카탈로그를 통해 이루어지기 때문에 오픈데이터 플랫폼의 활용에 있어서 표준 데이터 카탈로그의 지원은 중요한 요소중의 하나이다.

데이터 카탈로그 표준기술 중 대표적인 기술인 DCAT은 웹상에 존재하는 카탈로그 데이터간에 상호연동성을 제공하기 위한 W3C 표준으로 다양한 데이터 소스로부터 메타데이터를 읽어 데이터의 접근 및 활용을 가능하게 할 수 있도록 RDF 형태로 정의된다.

DCAT은 유연한 확장성으로 인해 CKAN, DKAN, Socrata와 같은 많은 오픈데이터 플랫폼에서 적용하고 있고 실제 data.gov와 data.gov.uk와 같은 많은 공공데이터 포털에서 데이터 연동을 위해 활용되고 있다. DCAT의 데이터 모델은 아래 그림 1과 같이 3개의 주 클래스인 Catalog, DataSet, Distribution 클래스와 1개의 중요 클래스인 Catalog Record 클래스를 중심으로 구성된다.

Catalog 클래스는 데이터셋의 집합을 표현하는 메타데이터들의 집합으로 데이터 카테고리를 의미한다. 따라서 Catalog 클래스는 Dataset 클래스들과 CatalogRecord클래스들을 포함한다.

Dataset 클래스는 실제 사용자에게 배포되는 실제 데이터셋의 집합을 의미한다. 따라서 사용자가 다운로드 받고자 하는 데이터셋의 이름, 지역, 발행일자등을 포함한다.

Distribution 클래스는 데이터셋이 사용자에게 제공되어지는 형태(format) 정보를 관리한다. 예를 들어 동일한 데이터셋이라도 사용자에게 CSV 파일, API, RSS 피드형태로 제공될 수 있다.

DCAT-AP[10]는 DCAT을 기반으로 유럽의 데이터 포털이 공공데이터를 발행 및 연동을 위해 정의한 DCAT 응용 프로파일이다. CKAN의 DCAT 지원은 실제로 DCAT-AP 지원을 의미하고 있다.

DCAT의 대표적인 활용 사례로, CKAN을 개발한 OKFNLab이 유럽 데이터 포털인 publicdata.eu 프로젝트를 통해 DCAT을 기반으로 CKAN 카탈로그를 확장하고 있으며 데이터 상호운용성을 지원하기 위한 표준화 작업을 진행중에 있다. 또한, CKAN은 DCAT을 지원하기 위해 RDF기반의 직렬화 기능을 포함하는 플러그인을 개발하여 추가함으로써 CKAN이 외부의 메타데이터와 상호운용성을 가지도록 하였다.

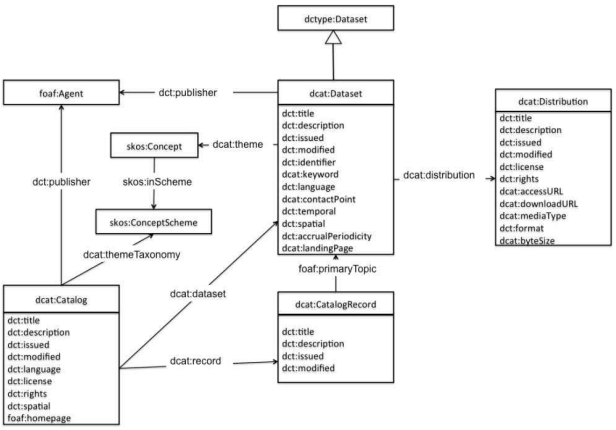

그림 1. DCAT 데이터 모델
Fig. 1. DCAT Data Model

Ⅲ. DCAT 기반 메타데이터 변환

본 장에서는 메타데이터 변환도구를 이용하여 데이터 유통 플랫폼의 데이터셋 정보를 DCAT 기반의 메타데이터로 변환하는 과정에 대해 기술한다.

그림2는 DCAT 변환도구를 이용한 메타데이터 변환 과정을 보여준다. 메타데이터 변환 과정은 크게 3단계로 구분할 수 있다. 첫 번째 단계에서는 데이터셋 정보를 얻기위한 데이터스토어 접근이다. 이 단계에서는 데이터베이스의 스키마 정보와 데이터셋 정보를 읽어온다. 두 번째 단계는 매핑 단계로 읽어온 데이터베이스 스키마와 DCAT간의 매핑 작업을 수행한다. 매핑 작업이 완료되면 마지막 단계로 DCAT RDF 파일생성을 수행한다.

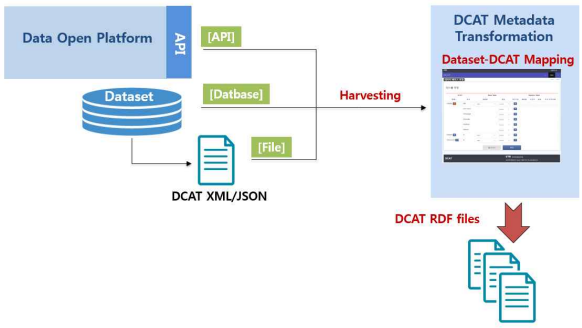

그림 2. DCAT 기반 메타데이터 변환
Fig. 2. DCAT-based Meatadata Transformation

3-1 데이터셋 정보 수집

메타데이터 변환도구는 데이터셋 정보를 수집하기 위해 해당 플랫폼의 데이터셋을 저장관리하는 데이터스토어에 접근해야 한다.

메타데이터 변환을 위해서는 해당 데이터 플랫폼의 데이터스토어에 접근하여 메타데이터의 스키마 구조와 메타데이터 정보를 검색할 수 있어야 한다. 메타데이터의 스키마 구조는 DCAT 모델과 매핑하는데 사용되고 데이터셋정보는 매핑후 변경된 스키마에 저장될 정보이다.

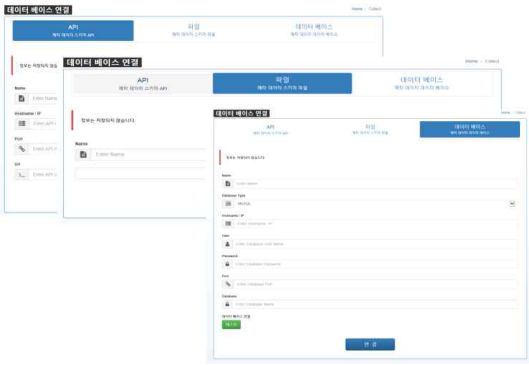

그림 3. 데이터스토어 접근을 위한 사용자 인터페이스
Fig. 3. User Interface for accessing Datastore

본 논문에서 제안하는 메타데이터 변환도구는 아래와 같이 같이 3가지 방법으로 데이터스토에 접근한다.

• 데이터베이스 계정을 이용한 데이터스토어 접근

해당 데이터 플랫폼의 관리자는 데이터베이스 계정을 이용하여 접근할 수 있다. 데이터베이스 계정을 이용하여 접근하는 경우는 데이터스토어의 정보를 다른 플랫폼과 공유하고 싶지 않을 경우 사용할 수 있는 방법이다.

따라서 이와 같은 경우는 데이터 플랫폼 관리자가 데이터베이스에 접근한 후 다음 단계인 스키마 매핑 작업을 직접 해야 한다.

• 오픈 API를 이용한 데이터스토어 접근

오픈 API를 이용하는 방법은 데이터스토어 정보를 제공하는 가장 일반적인 방법으로 해당 데이터 플랫폼은 사용자들에게 오픈 API를 제공함으로 사용자는 자유롭게 데이터스토어의 정보를 활용할 수 있게 된다.

오픈 API를 사용할 때 고려해야 할 점은 오픈 데이터 플랫폼마다 서로 다른 API를 제공한다는 점이다. 따라서 공통 오픈 API를 제공하여 상호연동을 원하는 데이터 플랫폼들이 공통 오픈 API를 지원하게 함으로써 플랫폼간 상호연동을 용이하게 하였다.

정의된 공통 오픈 API는 다음과 같다.

표 1. 데이터스토어 접근을 위한 오픈 API
Table 1. Open API for Datastore Access

HTTP Method	HTTP URL	기 능
GET	/dcat/tables	모든 테이블 정보 조회
GET	/dcat/tables/tb_members	특정 테이블(tb_members) 정보 조회
GET	/dcat/tables/tb_members?page=3&limit=100	페이지 단위의 테이블 정보 조회
GET	/dcat/tables/tb_members?fields=id,name,age	테이블 컬럼 조회
GET	/dcat/tables/tb_members?name=cho&city=seoul	테이블 조건 검색
GET	/dcat/tables/tb_members/name/tb_users/f_name? fields="tb_members.id", "tb_members.name", "tb_users.city"	테이블 조인 검색
GET	/dcat/tables/tb_members/name/tb_users/f_name?cond="tb_members.age>30" & fields-="tb_memebrs.id", "tb_members.name", "tb_users.city"	테이블 조인 검색

• 파일을 이용한 데이터스토어 접근

데이터 플랫폼이 API를 제공하지 못하거나 다른 플랫폼으로부터 데이터스토어의 접근을 허락하지 않을 경우에는 메타데이터를 공유하기 위한 방법으로 파일을 이용하여 데이터스토어의 정보를 제공하는 방법이 있다. 플랫폼 관리자는 공유가 가능한 스키마 및 메타데이터 정보를 XML 또는 json 형태의 파일로 생성하여 제공할 수 있다.

3-2 데이터셋-DCAT 매핑

메타데이터를 DCAT 기반의 메타데이터로 변환하기 위해서는 데이터셋과 DCAT 간의 매핑 작업이 이루어져야 한다. 그림은 데이터셋과 DCAT 매핑을 수행하는 화면을 보여준다. 그림에서 좌측은 DCAT 클래스와 속성 필드를 보여주고 우측은 데

이터셋을 저장 관리하는 스키마와 릴레이션 정보를 보여준다.

데이터셋과 DCAT간의 매핑시 모든 DCAT 필드를 데이터셋과 매핑할 필요는 없으며 데이터셋과 관련이 있는 경우에만 해당 필드를 매핑한다. 매핑을 하기 위해서는 우선 좌측의 DCAT 패널에서 매핑하고자 하는 DCAT의 클래스와 해당 클래스의 속성을 선택한다. 그리고 우측패널에서 해당되는 데이터베이스 테이블을 선택한다. 해당되는 테이블이 선택되면 컬럼 패널에 해당 테이블의 컬럼 목록들이 나타나고 이 컬럼들중 해당 컬럼을 선택한다.

기본적으로 이와 같이 DCAT 클래스와 속성을 데이터베이스의 테이블과 컬럼으로 매핑하는데 실제로 테이블내 컬럼이 데이터를 저장하고 있는 경우도 있지만 아이디 값을 가지고 있어 릴레이션을 구성하고 있거나 코드값을 가지고 있는 경우가 있다.

따라서 이런 경우에는 조인 연산을 통해 해당 값을 검색하여 가져와야 한다. 이러한 기능을 하는 것이 좌측의 릴레이션 패널이다.

만약 2개의 테이블이 릴레이션을 가지고 있고 한번의 조인 연산을 통해 데이터를 읽을 수 있다면 데이터베이스 패널에는 릴레이션을 갖는 테이블과 매핑되는 컬럼을 설정하고 릴레이션 패널에는 데이터베이스 패널에 해당되는 테이블과 릴레이션을 갖는 컬럼을 매핑한다. 마지막으로 뷰(view) 패널에는 우리가 검색하고자 하는 컬럼을 설정하여 해당 값을 읽어온다.

실제적으로 테이블간의 릴레이션은 1회 이상의 조인연산이 필요할 수도 있기 때문에 사용자 인터페이스에서는 테이블간 릴레이션을 1회 이상 가능하도록 제공하고 있다.

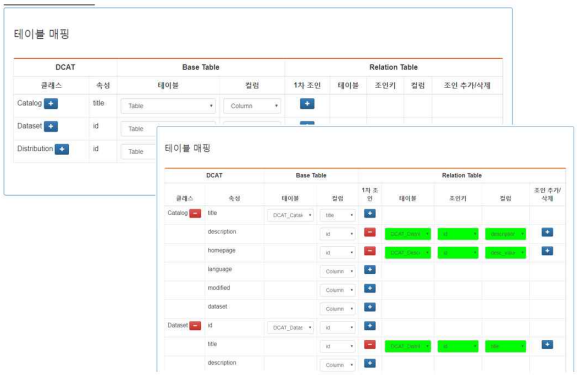

그림 4. 데이터셋-DCAT 매핑
Fig. 4. Dataset-DCAT Mapping

3-3 DCAT RDF 파일 생성

데이터셋과 DCAT간의 매핑작업이 끝나면 그림과 같이 매핑정보 요약을 통해 매핑 작업을 확인하고 오류를 수정할 수 있다. 데이터셋과 DCAT간의 매핑정보는 테이블 형태로 제공되고 가독성을 높이기 위해 그래프 형태로도 제공된다.

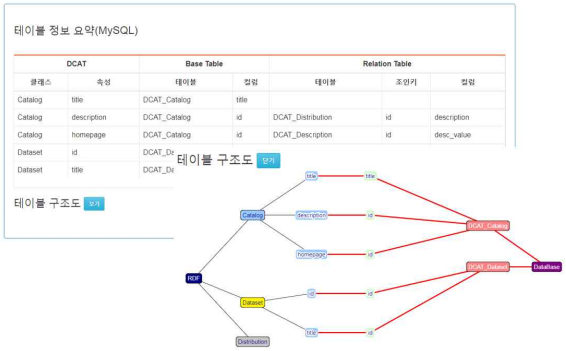

그림 5. 데이터셋-DCAT 매핑정보 요약
Fig. 5. Dataset-DCAT Mapping Summary

데이터셋과 DCAT간의 매핑작업이 완료되면 매핑 정보를 기준으로 데이터셋을 DCAT으로 변환한다. 이때 데이터셋의 일부가 DCAT으로 매핑되기 때문에 DCAT 하베스팅의 대상이 되는 부분은 매핑된 데이터셋 정보로 한정된다.

데이터셋은 RDF 형태로 변환되는데 2가지 형태로 변환될 수 있다. 첫번째는 카탈로그 엔드포인트(Catalog Endpoint) 형태로 하나의 카탈로그에 다수의 데이터셋 정보를 출력하는 형태이다. 두번째는 데이터셋 엔드포인트(Dataset Endpoint) 형태로 각 데이터셋 정보를 하나의 RDF 파일로 출력하는 형태이다.

생성된 DCAT RDF 파일은 DCAT 변환도구의 저장소에 저장된다. 아래 그림은 생성된 DCAT RDF 파일 내용과 생성된 RDF 파일의 관리화면을 보여준다. DCAT 메타데이터 변환도구는 RDF 파일과 함께 사용자 아이디, RDF 파일명, 생성일, 파일 크기, 접근한 데이터베이스 종류와 같은 정보를 함께 관리한다. 이외에도 DCAT 메타데이터 변환도구는 RDF 파일을 배포하는 기능을 가진다. 따라서 배포 기능을 통해 해당 플랫폼에 DCAT RDF 파일 정보를 전달함으로써 해당 플랫폼에서 DCAT 하베스팅을 가능하도록 한다.

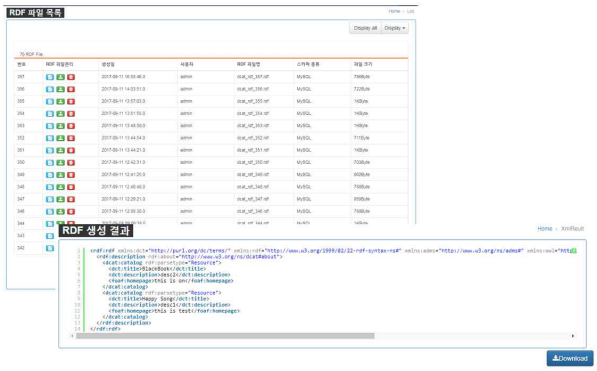

그림 6. DCAT RDF 파일 관리
Fig. 6. DCAT RDF File Management

Ⅳ. 결 론

전 세계적으로 정부기관을 중심으로 하는 공공데이터 개방이 이루어지고 있으며 주요 몇몇 국가들은 개방된 공공데이터의 다양성과 품질 및 사용자 활용 측면에서 상당한 수준에 이르렀다. 국내에서도 정부 3.0을 중심으로 공공데이터의 민간 개발에 대한 법률을 제정하고 로드맵을 작성하는 등 공공데이터의 활성화에 박차를 가하고 있다.

이와 같은 상황에서 데이터 포털간에 데이터셋을 공유하는 기능은 데이터 플랫폼이 가져야 할 반드시 필요한 기능중의 하나이다. 이에 본 논문에서는 기존의 플랫폼을 변경하지 않고도 플랫폼간 데이터 상호운용성을 지원할 수 있도록 DCAT 기반의 메타데이터 변환도구를 설계, 구현하였다.

DCAT 기반의 메타데이터 변환도구는 메타데이터의 변환을 통해 데이터 포털간 상호운용성을 가능하게 해 주기 때문에 DCAT을 지원하지 않는 많은 국내의 데이터 포털들이 추가적인 변경작업없이 용이하게 다른 데이터 포털들과 데이터셋을 공유할 수 있게 되었다.

실제로 DCAT 변환도구를 이용하여 국내의 공공기관에서 운영중인 데이터 포털중 한곳과 민간기업에서 운영하고 있는 데이터 포털 한곳을 대상으로 메타데이터 변환을 수행하였고 데이터 하베스팅을 수행하여 그 기능을 확인하였다. 현재는 국내 다른 데이터 포털들과도 메타데이터변환 도구 활용을 논의중이다.

하지만 국내 두 곳의 데이터 포털로부터 DCAT 메타데이터를 변환하여 하베스팅을 수행한 결과 몇가지 문제점을 발견하였다.

첫째, 데이터 포털들의 데이터셋 정보가 DCAT과 매핑되는 부분들이 많지 않아 변환되는 메타데이터의 정보량에 한계가 있었다. 특히 데이터셋 정보가 많지 않았던 민간 데이터 포털의 경우, 아주 기본적인 정보들만 변환되어 서비스를 하기위는 정보가 많이 부족하였다.

둘째, 메타데이터 변환 문제라기보다는 CKAN의 데이터 저장관리 문제로 CKAN은 DCAT으로 변환하여 저장할 때 중요 속성들을 제외한 나머지 속성들은 CKAN내에 추가 테이블(package_extra 테이블) 또는 추가 필드(resource 테이블의 extras 필드)에 <키, 값> 형태로 저장된다. 따라서 이러한 정보들을 활용해야 할 경우에는 데이터를 읽어 파싱을 해야 하기 때문에 성능에 많은 영향을 미치게 된다.

따라서 오픈데이터 플랫폼이 보다 향상된 상호운용성을 제공하기 위해서는 데이터 저장관리에 대한 연구가 필요하다. 또한 국내 데이터 포털의 요구사항을 분석하여 DCAT 모델을 확장 설계하는 것이 필요하다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2017-0-00253, 국제표준 기반 오픈 데이터 유통 플랫폼 확장 기술 개발)

참고문헌

[1] CKAN, The open source data portal software. Available: http://ckan.org/

[2] DKAN Open Data Platform. Availabel: https://getdkan.org/

[3] Socrata: Data-Driven Innovation of Government Programs. Available: https://socrata.com/

[4] W3C Data Catalog Vocabulary(DCAT). Available: https://www.w3.org/TR/vocab-dcat/

[5] DATA.GOV: The home of the U.S. Government's open data. Availble: https://www.data.gov/

[6] DATA.GOV.UK: Opening up Government. Availble: https://data.gov.uk/

[7] DATA.GO.KR: Korea Open Data Portal. Availble: https://www.data.go.kr/

[8] Datastore. Available: https://www.datastore.or.kr/intro.do

[9] Open Govenrment Platform (OGPL). Available: http://ogpl.github.io/index-en.html

[10] DCAT Application Profile for Data Portals in Europe. Available: https://joinup.ec.europa.eu/asset/dcat_application_profile/ description

[11] Open Data Platform and Open Data Strategy, NIA, IT& Future Strategy, 2013. No. 16

[12] S. Neumaier, J. Umbrich, A. Polleres, "Challenges of mapping current CKAN metadata to DCAT", W3C Smart Descriptions & Smarter Vocabularies (SDSVoc), 2016.

박경현(Kyounghyun Park)

1999년 : 충북대학교 컴퓨터공학과 (공학사)
2001년 : 충북대학교 전산학과 대학원 (이학석사)

2001년~현 재: 한국전자통신연구원 선임연구원
※관심분야 : 대용량 분산 데이터관리시스템, 빅데이터 플랫폼, 클라우드 컴퓨팅

원희선(Hee Sun Won)

1990년 : 연세대학교 전산학과 (이학사)
1992년 : KAIST 전산학과 대학원 (이학석사)
2016년 : KAIST 전산학과 대학원 (이학박사)

1992년~1999년: 한국방송(KBS) 기술연구소 연구원
2000년~현 재: 한국전자통신연구원 책임연구원
※관심분야 : 빅데이터, 클라우드 플랫폼, 오픈 데이터 거버넌스

류근호(Keun Ho Ryu)

1976년 : 숭실대학교 전산학과 (이학사)
1980년 : 연세대학교 대학원 전산전공 (공학석사)
1988년 : 연세대학교 대학원 전산전공 (공학박사)

1976년~1986년: 육군군수 지원사 전산실(ROTC 장교), 한국전자통신연구원(연구원), 한국방송통신대학교 전산학과(조교수)
1989년~1991년: Univ. of Arizona Research Staff
1986년~현 재: 충북대학교 소프트웨어학과 교수
※관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 유비쿼터스 컴퓨팅 및 스트림데이터 처리, 데이터 마이닝, 데이터베이스, 보안, 바이오 인포메틱스