DCAT 기반 메타데이터의 웹 출판을 위한 변환 기법

박진효*, 김기훈**, 김성희***, 윤주상***
*동의대학교 IT융합학과
**한국정보통신기술협회
***동의대학교 산업ICT기술공학
jhpark3679@gmail.com, channel@tta.or.kr, sh.kim@deu.ac.kr, jsyoun@deu.ac.kr

Transformation Method for Publishing DCAT based Metadata in Data Repository on Web

Jinhyo Park*, Kihun Kim**, Sung-Hee Kim***, Joosang Youn***
*Dept. of of IT Convergence, Dong-Eui University
**TTA(Telecommunications Technology Association)
***Dept. of Industrial ICT Engineering Engineering, Dong-Eui University

요 약

최근 데이터 산업 발전과 함께 데이터를 저장, 공유, 거래가 가능한 다양한 데이터 저장소와 거래소가 증가하고 있다. 대부분의 데이터 저장소 및 거래소는 데이터 검색과 공유를 위해 DCAT 기반 메타데이터를 구성하고 있다. 하지만 DCAT 기반 메타데이터는 웹 검색 엔진에서 검색이 잘되지 않는 문제점을 가지고 있다. 이는 웹에서 자원을 출판하기 위한 데이터 모델 기법이 Schema.org 방법을 사용하고 있기 때문이다. 본 논문에서는 이런 문제점을 해결하기 위해 DCAT 기반 메타데이터를 Schema.org 방법으로 변환할 수 있는 새로운 기법을 제안한다. 제안하는 변환 기법은 데이터 저장소와 거래소 내 데이터셋이 웹에서 잘 검색될 수 있는 웹 출판 기능을 지원한다.

1. 서론

최근 데이터를 활용하는 데이터 융합산업이 증가하면서 데이터 수집, 공유, 거래가 가능한 다양한 데이터 저장소와 거래소가 증가하고 있다. 이에 따라 데이터 저장소와 거래소 내에 존재하는 다양한 데이터를 효율적으로 검색할 수 있는 다양한 연구가 진행중이다. 특히, 데이터 저장소와 거래소 내에 저장된 데이터 정보를 공유할 수 있는 공통의 체계가 마련되어 있지 않아 데이터 저장소와 거래소 내에 존재하는 데이터를 공유하고 검색하는 데 어려움이 있다. 이는 데이터 저장소와 거래소에서 사용되는 메타데이터 체계가 통일되어 있지 않은 문제와 함께 메타데이터가 웹에서 잘 검색되지 않는 문제로 인해 발생한다.

이와 관련하여 RDA(Research Data Alliance)의 Research Metadata Schemas WG[1]에서 웹상에 구조화된 메타데이터를 출판하는 방법에 대한 표준 가이드라인이 개발 중이다. 웹은 Schema.org 방법을 통해 웹 자원에 대한 데이터 모델링 방법이 적용되고 있으며 이를 기반으로 검색 환경에서 자원을 검색하고 있다. 하지만, 데이터 저장소와 거래소 내에 메타데이터는 DCAT(Data Catalog Vocabulary) [2]을 기반으로 구성된다. 따라서 웹상에서 데이터 검색이 잘 이루어지고 있지 않아 데이터 검색에 대한 정확성이 매우 낮게 나타나고 있다. 본 논문에서는 이와 같은 문제점을 해결하기 위해 데이터 저장소와 거래소 내 DCAT 기반 메타데이터를 웹에 출판하기 위한 Schema.org 변환 기법을 제안하고자 한다. 본문은 2장에서 관련 연구를 기술하고 3장에서 메타데이터 변환 기법을 제안하며 4장에서 결론을 맺는다.

2. 관련 연구

DCAT은 웹에 게시된 데이터 카탈로그 간의 상호 운용성을 쉽게 하도록 설계된 RDF 언어 표준으로, 2014년 표준화된 후 2020년에 2.0 버전이 개정되었다. DCAT 표준은 EU, 미국 등 대부분의 오픈 데이터 포털에서 구조화된 메타데이터를 구성하는 데 사용되고 있다. DCAT은 13개의 클래스가 있으며 그중 6개의 주요 클래스가 있다. 주요 클래스는 표 1과 같다. [2, 3]

<표 1> DCAT 주요 클래스 [2]

Class	Discription
dcat:Catalog	개별 항목이 일부 리소스를 설명하는 메타데이터 레코드
dcat:Resource	카달로그의 메타데이터 레코드로 설명할 수 있는 데이터셋, 데이터 서비스 또는 기타 리소스
dcat:Dataset	데이터셋, 단일 에이전트가 게시하거나 선별한 데이터 모음
dcat:Distribution	다운로드 가능한 파일과 같은 액세스 가능한 데이터 세트 형식
dcat:DataService	하나 이상의 데이터셋 또는 데이터 처리 기능에 대한 액세스를 제공하는 인터페이스를 통해 액세스할 수 있는 작업 모음
dcat:CatalogRecord	정보와 관련된 카탈로그의 메타데이터 항목

대부분의 오픈 데이터 플랫폼 환경에서 메타데이터가 DCAT 주요 클래스를 기반으로 메타데이터가 구조화되어 있다. 실제로 대표적인 웹 검색 엔진인 Google 검색사이트에서 오픈 데이터 플랫폼 내에 데이터를 검색할 경우 데이터 검색 정확도가 매우 낮게 나오는 결과를 확인할 수 있다[4]. 이는 위에서 언급했던 것처럼 구조화된 DCAT 기반 메타데이터를 웹 검색에서 정확한 데이터 자원을 검색하지 못하고 있기 때문이다.

그림 1은 데이터 저장소와 거래소 내 메타데이터를 웹에 출판하여 검색 환경에서 데이터를 검색할 수 있는 과정을 보여준다. 메타데이터 출판자는 데이터를 Schema.org로 변환한 후, Schema.org 어휘를 사용해 마크업 메타데이터를 생성하고 랜딩 페이지에 포함시킨다. 랜딩페이지 URL을 사이트맵에 포함 시켜 잠재적인 다운스트림 소비자와 사이트맵을 등록한다. 데이터 에그리게이터는 크롤을 보내 사이트맵에서 URL을 가져와 랜딩페이지에서 구조화된 데이터를 노출 시킨 후 인덱스를 검색할 수 있게 한다.

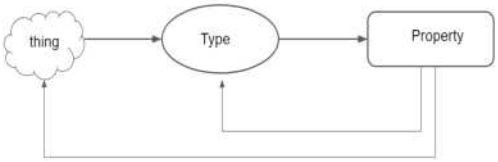
(그림 1) 데이터 출판 프로세스[1]

<표 2> DCAT 기반의 메타데이터를 Schema.org 데이터 모델로 변환 테이블

DCAT[2]		Schema.org[6]		Discription[2]
Class	Property	Property	Expected Type	
dcat: Dataset	dcat.distribution	Property::abstract	Text	데이터셋에 대한 설명
	dct:accrualPeriodicity	Property::creativeWorkStatus	DefinedTerm, Text	데이터셋의 발행하는 빈도
	dct:spatial	Property::spatial	Place	데이터셋이 적용되는 지리적 영역
	dcat.spatialResolutionlnMeter	Property::spatial	Place	미터 단위로 측정한 데이터셋에서 확인할 수 있는 최소 공간 해상도
	dct:temporal	Property::temporalCoverage	DateTime, Text, URL	데이터셋이 다루는 기간(시작 ~ 종료)
	dcat:temporalResolution	Property::datasetTimeInterval	Dataset	데이터셋에서 확인 가능한 최소 기간
	pxrov:wasGeneratedBy	Property::itemOffered	AggregateOffer, CreativeWork, Event, MenuItem, Product, Service, Trip	데이터셋을 생성하기 위해 비즈니스 컨텍스트를 생성하거나 제공하는 활동
dcat: Resource	dct:accessRights	Property::maintainer	Organization, Person	리소스에 액세스할 수 있는 사람에 대한 정보
	dct:conformsTo	Property::accessibilityFeature	Text	설명된 리소스가 준수하는 확립된 표준
	dcat:contactPoint	Property::contactPoints	Organization, Person	카탈로그된 리소스에 대한 연락처(vCard 사용)
	dct:creator	Property::creator	Organization, Person	리소스 생산을 담당하는 엔터티
	dct:description	Property::description	Text	개체에 대한 설명
	dct:title	Property::name	Text	개체의 이름
	dct:issued	Property::sdDatePublished	Date	개체의 공식 발행일
	dct:modified	Property::dateModified	Date, DateTime	개체의 변경, 업데이트 또는 수정된 가장 최근 날짜
	dct:language	Property::inLanguage	Language, Text	개체의 언어
	dct:publisher	Property::publisher	Organization, Person	개체을 사용할 수 있도록 하는 책임이 있는 엔터티
	dct:identifier	Property::identifier	PropertyValue, Text, URL	개체의 고유 식별자
	dcat:theme	Property::category	PhysicalActivityCategory, Text, Thing, URL	리소스의 기본 테마
	dct:type	Property::about	Thing	리소스의 특성 또는 장르
	dct:relation	Property::dataset	Dataset	카탈로그된 항목에 대해 지정되지 않은 관계가 있는 리소스
	dcat:qualifiedRelation	Property::maintainer	Organization, Person	다른 리소스와의 관계에 대한 설명/링크
	dcat:keyword	Property::keywords	DefinedTerm, Text, URL	리소스를 설명하는 키워드 또는 태그
	dcat:landingPage	Property::publishingPrinciples	CreativeWork, Organization, Person	카탈로그, 데이터 세트, 배포 및/또는 추가 정보에 액세스하기 위해 웹 브라우저에서 탐색할 수 있는 웹 페이지
	prov:qualifiedAttribution	Property::additionalType	URL	자원에 대한 어떤 형태의 책임이 있는 에이전트에 대한 링크
	dct:license	Property::copyrightNotice	Text	리소스를 사용할 수 있는 법적 문서
	dct:rights	Property::copyrightHolder	Organization, Person	저작권 표시와 같이 dct:license 또는 dct:accessRights 로 처리되지 않은 모든 권리와 관련된 설명
	odrl:hasPolicy	Property::publishingPrinciples	CreativeWork, URL	리소스와 관련된 권한을 표현하는 ODRL 준수 정책
	dct:isReferencedBy	Property::citation	CreativeWork, Text	카탈로그 리소스를 참조하거나 인용하거나 가리키는 발행물과 같은 관련 리소스

3. 메티데이터 출판을 위한 변환 기법

데이터 저장소가 메타데이터 레코드를 웹에 출판하려면 Schema.org 기반의 데이터 모델[5] 기법을 따라야 한다. 그림 2는 Schema.org 데이터 모델을 도시하고 있다. Schema.org 데이터 모델은 기존의 RDF 데이터 모델의 일부와 그 데이터 모델의 엄격한 규칙을 단순화하면서 웹 자원을 설명하는 장점이 있다. RDF, JSON-LD 등으로 직렬화될 수 있어 자원 항목의 유형과 속성을 쉽게 웹에 노출 시킬 수 있다.



(그림 2) Schema.org 데이터 모델[1]

표 2는 메타데이터 구성을 위한 DCAT 모델을 Schema.org 데이터 모델로 변환할 수 있는 변환 방법을 제시하고 있다. 특히, 표 2에 보이는 것처럼 DCAT의 'dcat:Dataset' 클래스와 'dcat:Resource' 클래스의 속성을 Schema.org의 속성으로 매핑할 수 있는 정보를 제시하였다.

메타데이터 변환 예를 살펴보면 'dcat:Dataset' 클래스에서 'dcat.distribution'은 Schema.org의 'Property::abstract' 속성으로 변환할 수 있다. 또한, 'dcat:Resource' 클래스의 'dcat:keyword' 속성은 Schema.org의 'Property::keywords' 속성으로 변환되며 이는 자원을 설명하는 키워드를 나타낸다. 이런 형식으로 DCAT 기반 메타데이터를 Schema.org 모델로 변환할 수 있다.

4. 결론

본 논문에서는 웹 검색 환경에서 데이터 저장소 및 거래소 내에 메타데이터를 웹에 출판하는 과정에서 구조화된 DCAT 메타데이터를 Schema.org 데이터 모델로 변환시키는 기법을 제안했다. 추후 표 1에 제시된 변환 테이블을 기반으로 변환 소프트웨어를 구현 후 성능을 테스트할 예정이다.

Acknowledgement

이 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2017-0-00061, 국내 ICT 표준 제개정 연구)

참고문헌

[1] M. Wu et.al. "Guidelines for publishing structured metadata on the Web V3.0," Research Data Alliance, Jun. 15, 2021. Accessed: Oct. 8 2021 [Online]. Available: https://www.rd-alliance.org/group/research-metadata-schemas-wg/outcomes/guidelines-publishing-structured-metadata-web-v30

[2] "Data Catalog Vocabulary (DCAT) - Version 2," W3C, last modified Feb 04, 2021, accessed Oct 06, 2021, https://www.w3.org/TR/vocab-dcat-2/.

[3] 정의현, 데이터 유통 및 데이터 품질 기술동향, 한국지능정보사회진흥원, 2021-5호, 7p, 2021

[4] "구조화된 데이터 작동 방식 이해", Google 검색 센터, last modified Aug 31, 2021, https://developers.google.com/search/docs/advanced/structured-data/intro-structured-data?hl=ko

[5] "Data Model," Schema.org, last modified Jul 07, 2021, accessed Oct 06, 2021, https://schema.org/docs/datamodel.html.

[6] "Full Hierarchy." Schema.org. last modified Jul 07, 2021, accessed Oct 06, 2021, https://schema.org/docs/full.html.