대구시 공공 링크드 데이터 구축 사례

정은미, 이용주
경북대학교 IT대학 컴퓨터학부
e-mail:jeunmi021@gmail.com, yongju@knu.ac.kr

A Case Study of Building Deagu Public Linked Data

Eunmi Jung, Youngju Lee
School of Computer Information, Kyungpook National University

요 약

우리에게 친숙한 웹은 1989년 팀 버너스리에 의해 시작되었지만 오늘날 웹이 없는 세상은 상상하기 어렵다. 이러한 웹의 발달로 많은 양의 데이터를 손쉽게 접할 수는 있게 되었지만 이제는 의미 있는 정보를 이끌어내는 것이 중요하게 되었다. 본 연구에서는 각각의 자원들이 연결된 데이터 중심의 웹을 구성하기 위해 링크드 데이터를 효율적으로 구축하기 위한 7단계 구축 방법을 제시하고, 본 구축 방법에 따라 시범적으로 실제 대구시 공공데이터를 이용하여 링크드 데이터를 구축해 본다.

1. 서론

우리에게 친숙한 웹은 전세계 대학과 연구소에 흩어진 물리학자들 간의 공동연구에 필요한 즉각적 정보교환 방안으로 1989년 팀 버너스리(Tim Berners-Lee)에 의해 시작되었지만 오늘날 웹이 없는 세상은 상상하기 힘들다[1]. 이러한 웹의 발달로 어마어마한 양의 데이터를 손쉽게 접할 수 있게 되었지만, 이제는 웹을 통한 의미 있는 데이터 정보를 이끌어내는 것이 중요하게 되었다.

링크드 데이터는 웹을 구성하는 데이터들 간의 연결을 목표로 기존의 문서 중심의 웹이 아닌 각각의 자원을 대상으로 상호 연결된 데이터 중심의 웹을 구성하는 것이다. 링크드 데이터는 사실 데이터를 포함하는 데이터 객체에 URI를 부여하고 이를 웹 프로토콜인 HTTP를 통해 발행하여 누구나 웹상에서 자유롭게 데이터를 활용할 수 있게 하는 기술이라 할 수 있다[2]. 즉 링크드 데이터를 통해 기존 문서 위주의 World Wide Web 전달 방식을 페이지가 아닌 데이터 간 연결 중심으로 전환하여 보다 풍부한 자원의 생산 및 효율적인 활용이 가능한 방식으로 웹을 지능화시키는 것이다[3].

본 논문에서는 각각의 자원들이 연결된 데이터 중심의 웹을 실제로 구성해 보기 위해 대구시 공공 데이터를 중심으로 구조화된 데이터의 출판과 연결을 위한 링크드 데이터 구축 방법을 제안하고, 시범 데이터를 선정하여 실제 링크드 데이터를 구축한다.

2. 링크드 데이터의 개요

위키피디아에 의하면, 링크드 데이터는 웹상에 존재하는 데이터를 개별 URI로 식별하고, 각 URI에 대한 링크 정보를 부여함으로써 상호 연결된 웹을 지향하는 모델로 정의한다. 이러한 링크드 데이터를 통해 사람이 이해하고 활용하는 문서 중심의 웹을 기계 또한 이해하고 자동으로 처리할 수 있는 데이터 중심의 웹을 구축할 수 있다[4]. 기술적으로 링크드 데이터의 핵심 아이디어는 HTTP URI의 사용이다. URI는 웹 문서들을 식별하는 것뿐만 아니라 임의의 실세계 객체들을 식별할 수 있다. 링크드 데이터를 구축하기 위해서는 비구조적인 데이터를 구조화하는데 시맨틱 형태로 표현해야 한다. 여기서 자원을 구조화한다는 것은 데이터를 RDF 형태로 표현하는 것이다[5].

RDF(Resource Description framework)[6]는 웹 상의 데이터를 교환하기 위한 표준 모델로서, 주어(subject), 서술어(predicate), 목적어(object)의 트리플(triple) 구조로 정보를 표현한다. RDF는 사람이 쉽게 읽고 이해할 수 있을 뿐만 아니라 기계적인 처리, 즉 응용프로그램들이 웹에 표현된 정보들을 처리하기 용이하여, 다양한 어플리케이션 영역에서 사용될 수 있다[7].

3. 국내 링크드 데이터 구축 사례

3.1 공공부문

공공데이터 포털(data.go.kr)에서는 국가서지, 한국사, 산업재산권, NDSL 등 10개의 LOD(Linked Open Data) 서비스를 제공하고 있다.

1) 국가서지

국립중앙도서관이 국가 대표도서관으로써 관리, 보존하고 있는 서지정보와 주제명, 저자명 데이터에 대해, 기존의 MARC 형태나 DBMS 형태의 데이터를 RDF 형식으

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1B02008553).

로 변환하여 LOD를 구축하였다.

2) 한국사

국사편찬위원회 인물 온톨로지와 시소러스, 한국학중앙연구원 민족문화대백과사전, 문화재청 유물/유적데이터와 각 유관 기관에서 보유 하고 있는 역사데이터를 융합하여 한국사에 관한 다양한 서비스나 어플리케이션 개발에 활용할 수 있도록 LOD 서비스를 제공한다.

3) 산업재산권

특허, 상표, 디자인 등 산업재산권 정보와 심판정보 등 행정정보, 분류코드 정보, 유의어 사전 등 5종에 대한 데이터 서비스를 제공한다. 기존 키워드 검색 중심에서 산업재산권 LOD를 통해 데이터 간의 연결성을 강화하여 보다 정확하고 정제된 정보를 사용자에게 전달할 수 있다.

3.2 민간부문

1) 성경 온톨로지

서울대학교 Linked Data Center에서 성경 속의 역사적 사건, 역사적 인물, 시대 등 성경에 대한 링크드 데이터를 구축한 사례이다. 공개 데이터를 신학교 교과과정과 연계하여 활용이 가능하다.

2) KDATA

한국과 관련하여 공개되어 있는 다양한 Linked Data 수집과 활용성이 높을 것으로 예상되는 새로운 링크드 데이터 구축 및 활용 서비스를 제공하며, 데이터 발행 신청을 하면 온톨로지 설계(모델링)-온톨로지(트리플)변환-Linked Data 발행 과정을 거쳐 무료로 데이터를 발행한다. 서울시 장난감 도서관, 대한민국 작가, 의료기관 등 다양한 데이터를 구축해왔다.

3) LOD 데이터 허브

“대한민국, 잇다!”라는 주제로 데이터를 쉽고 효과적으로 활용할 수 있는 기반 구축을 목표로 하고 있다. 다양한 영역의 데이터를 서로 연결하고, 데이터 매쉬업을 실현하기 위한 인프라를 구축하여 데이터의 재사용성을 향상시킨다. 구축데이터로는 행정구역, 도로명 코드, 지하철 노선, 지하철 역 등으로 기반 데이터와 분야별 데이터를 구축 범위로 선정하여 구축하였다[8].

4. 링크드 데이터 구축 방법

일반적으로 링크드 데이터를 구축하기 위해서는 데이터를 기계가 이해할 수 있도록 RDF 어휘(vocabulary)를 이용하여 RDF 데이터로 표현하고 링크를 생성해야 한다. 데이터가 링크드 데이터 형식으로 표현되어 있으면 사용자들은 자신이 찾은 정보에 링크되어 있는 부가적인 정보를 얻을 수 있다. 링크드 데이터 구축 단계는 다음과 같이 구성될 수 있다.

1) 데이터 특성 및 구조 분석

데이터를 발행하기 전에 구축 대상 자원들의 특성을 분석하고 데이터 구조를 분석하며, 타 시스템 구축 사례 및 링크드 오픈 데이터로의 구축 가능성을 조사한다. 데이터 모델, 메타데이터, 데이터 자원을 깊이 들여다보고 분석하는 단계를 통해 개괄적인 시스템 구성과 주요 클래스 관계도를 도출한다.

2) 추가적인 데이터 정제 작업

Database, XML, CSV 등과 같이 다양한 포맷과 분산된 데이터 소스로부터 획득한 데이터를 쉽고 효율적으로 데이터를 구축하기 위하여 추가적인 데이터 정제 작업이 필요하다[9]. 이 작업에는 데이터를 공개하지 않을 데이터를 제거하는 작업도 포함한다.

3) 링크드 데이터 모델링

발행하고자 하는 데이터를 RDF로 변환되도록 기존의 어휘나 새로운 모델을 선택해서 모델링을 수행한다. RDF는 주어-술어-목적어로 구성된 트리플을 기본 단위로 하고 트리플들의 집합인 그래프에 의해 리소스를 기술한다. RDF에서 사용하는 어휘는 RDF 스키마에서 정의되고, 개념이나 리소스 관계를 나타내거나 추론 가능한 논리를 기술하기 위해 온톨로지를 구축한다. 이러한 모델은 시간이 지남에 따라 변화될 수 있는 데이터의 영속성을 보장하기 위하여 유연한 동적 데이터 모델이 되도록 설계한다.

4) 기존 어휘 선택 및 독자적인 어휘 설계

RDF를 작성할 때에는 가능한 기존의 어휘를 사용하도록 한다. 공통적으로 인식된 어휘라면 상호 운용성을 높일 수 있고 많은 애플리케이션에서 공동 사용 가능하기 때문이다. 그러나 필요에 맞는 어휘가 없다면 내 데이터에 맞는 독자적인 어휘를 생성하도록 한다.

5) 저작권과 라이선스 검토

자원을 링크드 데이터로 오픈함에 있어 간과하지 말아야 하는 것이 저작권 및 라이선스 문제이다[10]. 발행하는 데이터가 광범위하고 효율적으로 재사용 가능하도록 라이선스를 지정하고 저작권을 보호해야 한다. 특히 기관의 내부 데이터를 다른 링크드 데이터 세트와 연계할 경우 다시 한 번 저작권과 라이선스 문제를 검토할 필요가 있다.

6) Non-RDF 데이터의 RDF 변환

기존에 비구조화된 형태로 구축된 Non-RDF 데이터(XML, Database, Spreadsheet, 등)를 RDF 형식으로 변환해 주어야 링크드 데이터 세트로 연계될 수 있다. 현재 매핑 솔루션으로 D2RQ가 오픈소스로 제공되고 있다. 국내의 경우 (주)탑쿼드란트코리아가 개발한 OntoBase 2.0이 이와 유사한 기능을 제공하고 있다.

7) RDF 데이터 간 링크

RDF 링크를 통해 전세계 데이터가 하나의 거대한 데이터베이스로 될 수 있고 이는 엄청난 부가가치가 발생할 잠재성을 지닌다. 이러한 연결은 최적화된 데이터 처리와 데이터 재사용을 위한 통합을 보장하고, 다른 도메인의 데이터와 연결되어 새로운 지식을 생성할 수 있게 된다[5].

5. 링크드 데이터 구축 사례

본 장에서는 대구시에서 제공하는 공공데이터를 활용하여 4장의 링크드 데이터 구축 방법에 따라 링크드 데이터를 실제로 한번 구축해본다. 서울시에서 제공하는 열린데이터광장(data.seoul.go.kr)의 링크드 데이터와 대구시에서 제공하는 공공데이터들이 유사하다고 판단됨에 따라 이번 연구에서는 서울시의 링크드 데이터를 상당 부분 참고하여 구축하였다.

1) 데이터 선정 및 수집

대구시에서 제공하는 여러 공공데이터들 중 활용도가 높은 데이터를 선정하고자 하였으며, 이에 따라 활용도가 높은 기준을 조회수로 두어 대구맛집 정보를 구축데이터로 선정하였다. 제공하는 데이터 형태는 Sheet(XLS, CSV, TXT), Open API 두 가지이지만, 데이터 분석 및 링크드 데이터 구축을 위해 본 연구에서는 Sheet 데이터를 사용하였다.

2) 데이터 분석 및 정제

대구맛집에 대해 제공하는 데이터 요소는 <표 1>과 같으며, 맛집에 대한 정보가 구체적으로 기술되어 있다. 구군분류와 업소명 외에 구체적인 주소에 대한 정보가 없는 점이 아쉬웠다.

<표 1> 대구맛집 정보 데이터 제공 항목

구군분류	음식분류	업소명
전화번호	영업시간	좌석수
주차장	홈페이지	가능 외국어
예약여부	유아시설	좌식유무
후식유무	메뉴	간단설명
지하철	버스	

3) 데이터 모델링

데이터 모델링을 위해 Schema.org, 서울 열린데이터 광장 LOD 온톨로지 스키마, KDATA 온톨로지를 재사용하고 참조하였다. 기존 어휘를 이용하여 나타낸 rdf 데이터를 Turtle 형태로 나타내면 다음과 같다.

```
@prefix schema: <http://schema.org/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix kdp: <http://data.kdata.kr/property/> .
@prefix seoul: <http://lod.seoul.go.kr/ontology/> .
<http://localhost/resource/daegufood/1>
    rdfs:label "우리식당" ;
    kdp:locatedIn "달서구" ;
    schema:telephone "053-000-0000" ;
    schema:openingHours "10:00 ~ 21:00" ;
    kdp:seatCount "30";
    schema:url "woori.test";
    schema:availableLanguage "영어";
    seoul:hasParkingLots "true"^^xsd:boolean;
    seoul:numberOfParkingLots "20";
    rdfs:description "푸짐한 양으로 유명한 한식전문점" .
```

4) 어휘 설계

이전 단계에서 활용이 가능한 어휘를 제외한 식당분류, 예약여부, 유아시설, 좌식유무, 후식유무 항목에 대한 어휘를 새로 설계하였다. 기존의 “좌석수” 항목의 경우 분리된 공간(방)에 대한 정보가 있어, 공간유무에 대한 항목을 추가하고 방의 수보단 유무에 초점을 맞추었다. 지하철은 지하철역, 출구번호, 출구와의 거리에 대한 정보, 버스는 버스번호가 아닌 버스정류장에 대한 정보로 데이터를 정제한 후 어휘를 설계하였다.

<표 2> 새로운 어휘 설계

hasSeparationSpace	분리된 공간 유무
	Datatype property(boolean)
menuInfo	메뉴
	Datatype property(string)
RestaurantType	식당분류
	Datatype property(string)
hasReservation	예약여부
	Datatype property(string)
isSedentary	좌식유무
	Datatype propert(/boolean)
hasDessert	후식유무
	Datatype property(boolean)
hasFacilitiesForChildren	유아시설유무
	Datatype property(boolean)
subwayStationInfo	지하철
	Datatype property(string)
busStopInfo	버스
	Datatype property(string)

새로이 설계한 어휘를 포함한 RDF 데이터는 다음과 같다.

```
@prefix schema: <http://schema.org/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix kdp: <http://data.kdata.kr/property/> .
@prefix seoul: <http://lod.seoul.go.kr/ontology/> .
@prefix ex: <example/> .
<http://localhost/resource/daegufood/1>
    rdfs:label "우리식당" ;
    kdp:locatedIn "달서구" ;
    schema:telephone "053-000-0000" ;
    schema:openingHours "10:00 ~ 21:00" ;
    kdp:seatCount "30";
    schema:url "woori.test";
    schema:availableLanguage "영어";
    seoul:hasParkingLots "true"^^xsd:boolean;
    seoul:numberOfParkingLots "20";
    rdfs:description "푸짐한 양으로 유명한 한식전문점" ;
    ex:RestaurantType "한식" ;
    ex:busStopInfo "시청" ;
    ex:hasDessert "true"^^xsd:boolean ;
    ex:hasFacilitiesForchildrren "false"^^xsd:boolean ;
    ex:hasReservation "불가" ;
    ex:hasSeparationSpace "false"^^xsd:boolean;
    ex:isSedentary "true"^^xsd:boolean ;
    ex:menuInfo "정식 4,000원" ;
    ex:subwayStationInfo "주변에 지하철이 없습니다." .
```

5) 저작권과 라이선스 검토

이번 연구에서 사용하는 공공데이터는 상업적 이용 및 콘텐츠 변경이 허용되어 본 연구에 맞게 데이터들을 정제하였다. 데이터 구축을 위해 사용된 온톨로지는 저작자 표시와 비영리 목적으로 공유 및 변경이 가능하였다.

6) RDF 변환

정제한 xls데이터를 MySQL로 데이터베이스를 구축하였고, 이를 RDF로 변환해주는 오픈소스 D2RQ를 활용하여 전체 데이터를 RDF 데이터로 변환하였다.

(그림 1) 링크드 데이터를 발행한 모습

7) RDF 데이터간 링크

이번 연구에서 구축한 데이터들은 시범 데이터로 실제 링크들과의 연계는 연구 범위에 포함시키지 않았다.

6. 결론

본 논문에서는 링크드 데이터를 구축하기 위한 방법 7단계: ① 데이터 특성 및 구조 분석, ② 추가적인 데이터 정제 작업, ③ 링크드 데이터 모델링, ④ 기존 어휘 선택 및 독자적인 어휘 설계, ⑤ 저작권과 라이센스 검토, ⑥ Non-RDF 데이터의 RDF 변환, 그리고 ⑦ RDF 데이터 간 링크를 제안하고 있다. 또한 제안한 구축 방법을 대구시에서 제공하는 공공데이터를 이용하여 시범 구축해보았다. 이번 연구에서는 극히 일부에 대해서 데이터를 구축하였으나 이를 바탕으로 앞으로 더 많은 데이터들이 실제로 구축되어 연계된다면 다양한 방면에서 활용도가 높을 것으로 기대된다.

참고문헌

[1] 이재호·양정진, “시맨틱 웹 : 차세대 지능형 웹 기술,” TTA저널 제81호, 2002, pp.79-85.

[2] 조명대·오원석·박진호, Linked Data 연구개발보고서: 주제명, 저자명 전거데이타 중심, 국립중앙도서관, 2011.

[3] 이용주, “링크드 데이터: 빅데이터 구축의 핵심 플랫폼,” 한국디지털경영학회 2014년 춘계학술발표대회논문집, 2014년 5월, pp. 237-244.

[4] 이병하 외 4인, 알기 쉬운 Linked Open Data, 한국정보화진흥원, 2015.

[5] 이용주, "링크드 데이터 구축 및 검색 기법," 한국정보처리학회 2014년 추계학술발표대회논문집, 제21권 제2호, 2014년 11월, pp. 1057-1060.

[6] https://www.w3.org/RDF/

[7] 황석형·조동헌, “형식개념분석법을 이용한 링크드 오픈 데이터 클라우드의 RDF데이터 분석,” 한국컴퓨터정보학회논문지, 제22권 제6호, 2017, pp. 57-68.

[8] 이병하·김택훈·박진호, 링크드 오픈 데이터 국내 구축 사례집, 한국정보화진흥원, 2014.

[9] F. Bauer and M. Kaltenböck, "Linked Open Data: The Essentials," A Quick Start Guide for Desicion Makers, 2012.

[10] 노영희, “dCollection의 링크드 데이터 구축에 관한 연구,” 한국도서관·정보학회지, 제43권 제2호, 2012, pp. 247-271.