

서울역사박물관 오픈데이터의 Linked Data 발행에 관한 연구

A study on Linked data publishing of Open data in Seoul museum of history

도솔기, 성균관대학교 대학원 문헌정보학과, sinhwask@gmail.com

한상은, 성균관대학교 대학원 문헌정보학과, silver86eun@gmail.com

Seulki Do, Dept. of Library and Information Science, Graduate School of Sunkyunkwan University

Sangeun Han, Dept. of Library and Information Science, Graduate School of Sunkyunkwan University

정부 및 기관, 개인에게 부가가치를 제공하는 공공 오픈데이터를 Linked Data로 발행하기 위한 다양한 시도들이 계속되고 있는 현 상황에서, 공공 오픈데이터인 '서울역사박물관의 유물 데이터'를 대상으로 데이터 정제 및 Linked Data로 발행하는 작업을 수행하여 발행 과정에서 나타나는 제약사항들에 대해 검토하였다. 이를 통해 정부 및 각 기관들, 개인이 데이터 발행자 및 이용자의 입장에서 공공 오픈데이터를 활용할 때 고려해야 할 사항들로 데이터 공개 시 데이터에 대한 명확한 설명 제시, 데이터 생애주기에 걸쳐 양질의 데이터 생산 및 공개, 데이터 발행자와 이용자 간의 지속적인 커뮤니케이션을 제언하였다.

1. 연구의 필요성 및 목적

오픈데이터는 웹상에서 누구든지 자유롭게 사용, 재사용, 재배포가 가능하도록 공개된 데이터로 2009년 팀 버너스 리(Tim Berners Lee)의 Ted 강연과 가트너(Gartner)의 '오픈 데이터가 빅데이터보다 더 많은 가치를 가져다 줄 수 있다'는 언급에 따라 중요성 및 관심이 증가하게 되었다.

이는 공공정보의 가치가 재조명 받는 계기가 되었다. 특히 인터넷 매체의 발달은 공공정보의 축적, 가공, 전달 작업을 용이하게 만들어 주었다. 정부는 공공정보 공개를 통해 시대적 요구에 대한 부응 및 새로운 공공서비스 개발의 토대로 사용할 수 있게 되었고, 공공 기관은 업무 부담을 경감시킬 수 있게 되었으며, 개인은 공공정보를 활용해 다양한 서비스를 제공받거나 새로운 부가가치를 스스로

만들어낼 수 있게 되었다.

현재 정부기관 및 유관기관들은 API, Text, XML, Excel spreadsheets 등의 형식으로 데이터를 공개하고자 노력하고 있다. 그러나 다양한 형식으로 제공되는 데이터는 재사용과 통합을 어렵게 한다. 따라서 오픈데이터는 공개 단계에서 그 목적에 맞게 다양한 출처의 신뢰할 수 있는 데이터를 통합하여 재사용할 수 있게 하고 상호운용성을 높여 효과적으로 활용할 수 있도록 해야 한다(김학래, 최재화, 2011).

미국과 영국 등의 국가에서는 오픈 공공데이터를 위에서 언급한 이점을 가진 링크드 데이터(Linked Data, 이하 LD)로 발행하고 있다.

LD는 다양한 형식으로 공개되어 있는 오픈 데이터를 의미적으로 연결할 수 있는 기반으로 RDF(Resource Description Framework) 트리플 형식으로 웹상에 발행하며, 각각의 자원은 URI라는 고유 식별자를 통해 연결되어

데이터의 웹(Web of Data)을 형성한다.

국내에서도 오픈데이터를 LD로 발행하는 연구가 활발하게 진행되었다. 문헌정보학 분야에서는 노영희(2012), 박옥남(2012) 등이 이용자의 관점에서 서지 정보들을 연결시키는 데 중점을 둔 연구를 진행하였고, 의학 분야에서는 서동민 외(2011) 연구자들이 이용자에게 혼란을 주는 용어들의 연결에 중점을 둔 LD 발행에 대해 연구하였으며, 공공 분야에서도 김학래, 최재화(2011), 이현정, 남영준(2012) 등이 효과적으로 정부데이터를 활용하기 위한 연구를 진행하였다.

이러한 연구들을 통해 공공 오픈데이터의 중요성이 커지고 있음을 확인할 수 있다. 그러나 기존의 연구들은 특정 도메인 분야에서 LD 발행 방법에 대해 기술하거나 국가 차원에서의 표준화 방안에 대해 제안하는 데 그치고 있다. 이에 공공 오픈데이터를 LD로 발행해보고 그 과정에서 나타나는 제약사항을 검토하여 LD 발행 시 고려해야 할 사항들에 대해 알아볼 필요가 있다.

따라서 본 연구에서는 공공 오픈데이터인 ‘서울역사박물관 유물 데이터’를 LD로 발행하는 작업을 수행하여 발행 과정에서 나타나는 어려움들을 확인해보고자 한다. 이를 통해 추후 데이터 발행자, 이용자 입장에서 공공 오픈데이터를 공개하고 활용하고자 할 때 고려해야 할 사항들에 대해 제안하고자 한다.

2. 오픈데이터의 LD 발행 사례

1) 서울 열린 데이터 광장

서울 열린 데이터 광장 Linked Data Beta 서비스는 공공데이터 활용과 공유 사례를 마련하고자 문화시설, 문화재를 중심으로 LD 기술을 구현하였다. 시맨틱 질의&검색, 시맨틱 관계 탐색, 패킷 네비게이션 등의 서비스를 제공하고 있다.

2) 제주도 Linking Open Data

제주도는 제주의 인문지리정보의 공개 및 연계 목적으로 인문지리정보 자료를 온톨로지 인스턴스화하여 제공하고 있다. 본 사이트는 이용자에게 인문지리 LOD를 제공하며 관광, 기후, 문화, 역사, 자원 등의 도메인에 대한 자료들을 LD로 발행하였다.

3) 미국의 data.gov

오바마 정부는 열린 정부 구현을 목표로 공공 정보 공개 전략을 수립하여 Data.gov 사이트를 통해 공공기관이 보유하고 있는 정보를 공개하고 있다. 공공데이터의 원활한 활용을 위해 RDF로 데이터를 표현하여 표준화했고, 이용자에게 RDF 트리플 및 데모 프로토타입을 제공하고 있다.

4) 영국의 data.gov.uk

영국은 data.gov.uk를 통해 미국 정부와 마찬가지로 정부데이터를 공개하고, 연관 서비스를 강화시키고 있으며 RDF 제공뿐만 아니라 SPARQL Endpoint를 직접 제공하고 있다.

이 외에도 지리정보를 LOD 형태로 제공하는 솔트룩스, 독일국립도서관, 미국의회도서관의 서지정보의 LOD화 등 각국에서 공공 오픈데이터를 LD로 발행하기 위한 다양한 사례를 확인할 수 있다. 이를 통해 공공 오픈데이터의 활용이 국가적인 차원에서 주요하게 진행되고 있다는 점을 알 수 있다.

3. 서울역사박물관 유물 데이터의 LD 발행

3.1 데이터 수집

서울 열린 데이터 광장에 공개된 데이터 중 Spreadsheets 형식으로 된 ‘서울역사박물관 유물데이터’를 이용하였다.

3.2 LD 발행 과정

본 연구에서는 온톨로지 모델링을 통해 RDF 저장소를 이용하여 LD를 발행하였다.

1) 데이터 정제

Open Refine 툴을 이용해 유물 데이터의 정제작업을 수행하였다. 이 과정에서 불용데이터 삭제 및 세분화가 가능한 데이터들을 세분화하는 작업을 진행하였다.

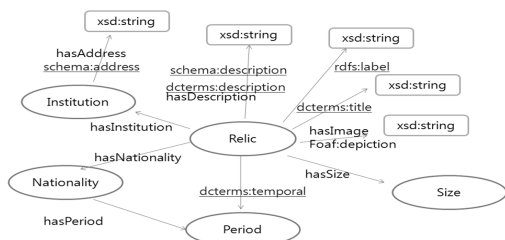
<표> 데이터 정제 예시

	정제 전	정제 후	건 수
괄호	홍씨추증(교지) (洪氏追贈(教旨))	홍씨추증, 교지, 洪氏追贈, 教旨,	164건
불용 데이터	백자대접(白磁大?)	백자대접	55건
특수 기호	"노안당"명 갱기 (老安堂銘羹器)	노안당 명 갱기	68건

2) Class, 관계 정의 및 어휘 선정

데이터의 내용 및 구조를 파악하여 Class와 Property들을 정의하였으며 DC(title,description), FOAF(depiction),Schema.org(address,description)등의 어휘를 선정하였다. 또한 DBpedia, Freebase와 같이 RDF dataset을 제공하는 곳과의 연결가능성을 검토하였다.

3) 위에서 정의한 내용들을 RDF 그래프 형식으로 표현하였다.



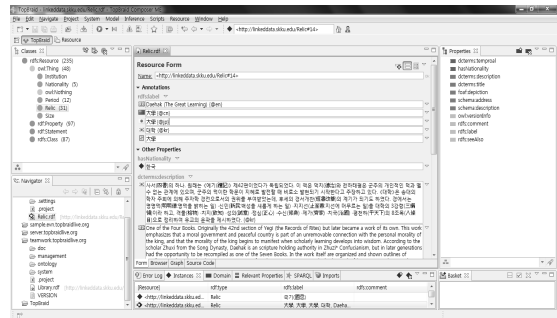
<그림 1> RDF 그래프 및 어휘선정 예

4) URI 패턴을 정의

URI 패턴은 LD의 중심사이트인 DBpedia를 기반으로 하였다.

5) 온톨로지 작성

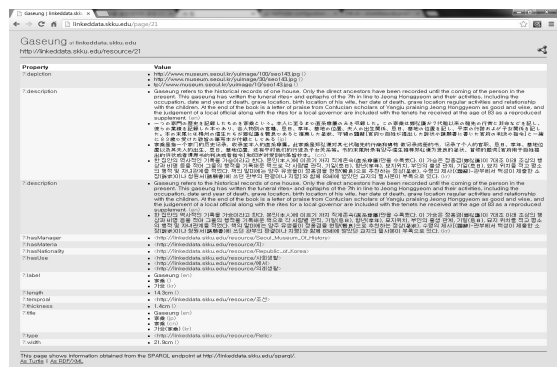
앞에서 작성한 RDF 그래프와 관계 및 어휘를 TBC(TopBraid Composer) 툴을 이용하여 온톨로지로 작성하였다. 또한 이 과정에서 작성한 온톨로지에 인스턴스를 입력하였다.



<그림 2> 온톨로지 작성 및 인스턴스 입력

6) 웹에 발행

RDF 저장소를 통해 웹에 발행하였다.



<그림 3> LD 발행 화면

4. 결론 및 제언

‘서울역사박물관 유물 데이터’를 LD로 발행하는 과정에서 나타난 주된 제약사항들은 다음과 같다.

데이터 정제 과정에서는 데이터 자체의 결함에 의한 어려움이 주를 이루었다. 데이터 깨짐 현상, 각기 다른 언어로 작성된 동일 데이터가 존재하는 등 이로 인해 데이터 통합의 과정을 개별적으로 거쳐야 했으며 데이터 정제 과정에서 많은 시간 및 노력을 필요로 하였다.

내·외부 링크 작업 과정에서는 한글로 된 데이터에 제약이 있어 원활한 연결 작업이 어려웠다. 이는 추후 외국인 이용자를 위해서도 영문화 작업이 이루어져야 할 것으로 보인다.

위의 제약사항들은 '서울역사박물관의 유물 데이터'에 국한되어 나타나는 것이기에 일반화 시키기에 한계가 있지만 발행자, 이용자 모두가 만족할 수 있는 오픈데이터의 활용을 위해서 몇 가지 사항들을 제안하고자 한다.

첫째, 데이터 공개 시 해당 데이터에 대한 명확한 설명이 수반되어 이용자가 용이하게 데이터에 대해 이해할 수 있도록 해주어야 한다.

둘째, 데이터를 보유, 관리, 제공하는 공공기관에서는 데이터 생애주기 전반에 걸쳐 양질의 데이터를 생산하고 공개해야 할 것이다.

셋째, 데이터를 공개하는 기관과 이용자 간의 원활한 커뮤니케이션 장이 마련되어야 할 것이다.

위에서 제안한 내용들은 데이터 발행자 및 이용자가 올바르게 데이터를 사용하여 이후 다른 이용자가 재사용할 수 있는 기회의 폭을 넓혀 오픈데이터의 의의를 지속시킬 수 있다는 데 의미가 있다.

2013년 7월 30일자로 공공기관이 보유·관리하는 데이터의 제공과 이용에 관한 사항을 규정한 '공공데이터의 제공 및 이용활성화에 관한 법률'이 공포되어 10월 31일자로 시행됨에 따라 공공데이터의 이용을 통한 가치창출이 증가하게 될 것이다. 따라서 본 연구에서 제안한 사항들은 추후 정부 및 각 기관에서 자관의 데이터를 오픈데이터로 공개할 시 참고자료로써 의미가 있을 것이다.

참고문헌

- 김학래, 최재화. 2011. 공개 정보데이터의 현황과 과제-시맨틱웹관점에서. 『e-비즈니스연구』, 12(3): 371-393.
- 노영희. 2012. dCollection의 링크드 데이터 구축에 관한 연구. 『한국도서관정보학회지』, 43(2): 247-271.
- 박옥남. 2012. 전거통제 기반 Linked Data 활용. 『한국기록관리학회 학술발표논문집』, 2012(5): 25-48.
- 서동민, 이승우, 이승복, 이상호, 성원경. 2011. 링크드 데이터를 활용한 의학용어 검색엔진 개발. 『한국인터넷정보학회 추계학술발표대회 논문집』, 2011(11): 81-82.
- 이현정, 남영준. 2012. 공공데이터베이스의 Linked Open Data 구축을 위한 가이드라인 설계. 『한국정보관리학회 학술대회논문집』, 2012(8): 63-68.
- Tom Heath, Christian Bizer. 2011. "Linked data: evolving the web into a global data space". CA: Morgan & Claypool.
- UK Cabinet Office. [cited 2013.08.02]. <http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf>.
- Tim berners-Lee Ted 강의. [cited 2013.07.31]. <http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html>.
- 공공데이터포털 [cited 2013.07.31]. <<http://data.go.kr>>.
- 미국 공공데이터 포털. [cited 2013.08.02]. <<http://www.data.gov>>.
- 서울 열린 데이터 광장 [cited 2013.07.31]. <<http://data.seoul.go.kr>>.
- 영국 공공데이터 포털. [cited 2013.07.31]. <<http://data.gov.uk>>.
- 제주도 LOD 시스템. [cited 2013.07.31]. <<http://lod.jeju.go.kr>>.