공공데이터베이스의 Linked Open Data구축을 위한 가이드라인 설계

A study on designing guidelines for Linked Open Data organization of national databases

이현정, 중앙대학교 대학원 문헌정보학과, caulis98@gmail.com
남영준, 중앙대학교 문헌정보학과, namyj@cau.ac.kr
Hyun-Jung Yi, Dept. of Library and Information Science, Graduate School of Chung-Ang University
Young-Joon Nam, Prof. of Library and Information Science, Chung-Ang University

공공데이터베이스는 공무상의 활용뿐만 아니라 민간의 창의성과 결합할 경우 새로운 비즈니스와 일자리를 창출할 수 있는 잠재력을 가지고 있다. 미국과 영국 등 해외 주요국들은 공공정보 재사용의 가치를 깨닫고 공공데이터를 Linked Data화하는 작업을 진행하고 있다. Linked Data는 웹에서 자유롭게 데이터를 개방하여 연계할 수 있도록 하는 네트워크 기술이다. 본 고에서는 국내 공공정보의 개방과 재사용을 지원하기 위한 방안으로 Linked Data 구축을 제안하고, 이를 위해 데이터, 시스템, 서비스의 3가지 측면에서 표준화 방안을 제안하였다. 데이터 표준화 측면에서는 데이터 표현 및 접근에 관한 표준을 준수하고, 표준화된 데이터모델과 데이터구조에 대한 정보와 함께 완전하게 수요자 입장에서 배포해야 한다. 특히 데이터의 상용화나 재가공에 아무런 제약이 없는 완전한 공개를 원칙으로 해야 한다. 시스템 표준화 측면에서 Linked Data 플랫폼은 트리플 스토어, 원데이터의 트리플 변환기, 그리고 추론기로 구성되어야 한다. 서비스 표준화 측면에서는 Linked Data를 이용자에게 다양한 포맷으로 제공할 수 있는 인터페이스가 제공되어야 한다. 무엇보다도 공공정보의 공개와 재사용성을 위한 국가적 차원의 거버넌스와 지원이 마련되어야 공공정보의 Linked Data 플랫폼이 온전히 이루어질 수 있을 것이다.

1. 서론

공공기관은 정책 수립과 같은 공적 기능을 수행하기 위해서 교통이나 물가, 범죄율 등의 정보가 필요하다. 때문에 많은 비용을 들여 여러 분야의 폭넓은 정보를 수집해 왔다. 이들 정보는 본래의 목적을 달성하게 되면, 그 수명을 다해 폐기되거나 방치되는 것이 일반적이었다. 그러나 인터넷과 디지털 기술, Open API와 매쉬업 등은 공공정보의 새로운 활용 가능성을 열어주었다. 정보의 축적과 가공, 전달이 용이해지면서, 공공정보가 지식정보의 원천으로 재조명받기 시작한 것이다. 우리나라에서도 2009년 고등학생이 공공기관 홈페이지에 공개된 정보를 이용하여 개발한 버스도착 앱이 폭발적인 인기를 끌면서, 그 활용 가능성과 수요를 확인한바 있다. 공공정보의 재사용은 바로 이 스마트폰 앱이 개발된 것처럼, 공공정보를 민간에 제공하여 새로운 상품과 서비스를 개발하는 기초자료로 이용하려는 것이다. 이로써 행정기관은 업무의 부담을 덜고 고품질의 행정서비스를 제공할 수 있게 되며, 국민은 다양하고 고도화된 상품과 서비스를 제공받을 수 있게 된다. 민간사업자는 자본없이도 창의성을 발현하여 커다란 수익을 창출할 수 있는 기회를 얻게 되며, 이는 산업진흥으로 이어져 국가경쟁력 강화로 이어지게 된다(최진원, 2012).

2. Linked Open Data

Linked Open Data (LOD)는 시멘틱웹이 표방하는 데이터웹(Data Web)을 구체적으로 구현하는 방법으로 인터넷 상의 각 사이트에서 RDF 형식의 데이터를 RESTful 프로토콜을 사용하여 정형화 데이터를 제공하는 것을 말한다.

Linked Data를 통해 관련된 데이터를 서로 연결함으로써 다음과 같은 장점을 얻을 수 있다.

- 데이터를 URI와 RDF, HTTP를 통해 연결하여 사용할 수 있으므로 내가 만든 데이터가 아니라도 Linked Data로 연결되면 하나의 지식베이스처럼 사용할 수 있다.
- Linked Data를 통해 공개된 데이터를 이용하면 자신이 원하는 데이터가 이미 존재하는지, 어디에 존재하는지 알 수 있으므로 시스템의 사일로(silo) 문제에 의해 발생된 불필요한 데이터 중복의 문제를 해결할 수 있다.
- 시멘틱웹 표준인 RDF 형태의 데이터로 발행하므로 마치 하나의 글로벌 데이터베이스처럼 질의하고 이용할 수 있으며, 이를 통해 상호운용성을 높이고, 데이터 통합을 용이하게 할 수 있다.
- URI로 구별되는 데이터 리소스의 자유로운 접근 및 이용이 가능하고, SPARQL Endpoint를 통해 SPARQL 질의가 가능하며, 이를 응용 프로그램 상에서 이용할 수 있으므로, OpenAPI에 비해 데이터 접근을 더욱 구체화할 수 있어 데이터 지향의 매쉬업을 할 수 있다.
- 초창기의 웹과 같이 데이터의 자유로운 연결과 이용은 새로운 데이터를 생산하고 양질의 데이터는 트래픽이 증가하게 되는 네트워크화(Network of Data, Cloud of Data)가 지속될 것이다. 또한 초기의 LOD Cloud에 진입한 데이터 셋들은 향후에 선점효과를 누릴 수 있을 것이다(한국문화정보센터, 2012).

Linked Data를 구축하기 위해서는 비구조적 데이터를 구조화하되 시맨틱 형태로 표현해야 한다. 자원을 구조화하는 것은 데이터를 RDF로 표현하는 것이며, 하나의 자원을 주어, 술어, 목적어로 표현하는 것이며, 각각의 자원은 URI를 통해 고유 식별자를 가지고 여러 가지 속성들을 이용하여자 자원들 간의 링크를 생산하는 것이다. Linked Data를 저작하거나 편집할 수 있는 시스템 또는 도구는 Loomp, Tabulator, mSpace Data Picker 등이 있다(노영희, 2012).

3. 해외 사례

3.1 미국

2009년 들어 공공 데이터의 공개운동이 본격화되고 있다. 연방 정부에서 생산된 기계 가독형 데이터셋을 공개함으로써 정부의 투명성을 높이고 부가가치를 창출하고자 하는 목적으로 오바마 정부에서 구축하였다. 연방 정부의 데이터에 대한 접근을 개선하고 이에 따라 혁신적인 아이디어를 장려함으로써 정부의 벽을 뛰어 넘어 데이터의 창조적 이용을 확대하고자 하였다. 구축된 data.gov의 일부를 RDF로 변환하여 Linked Data cloud에 합류하고자 하였으며, Tim Berners-Lee와 함께 시맨틱 웹의 창시자인 Jim Hendeler교수의 Tetherless World 팀이 프로젝트를 수행하였다. 현재까지 565개의 데이터셋을 RDF화 하였으며, 5,435,413,491개의 트리플이 구축되었다(http://data-gov.tw.rpi.edu/wiki/Data.gov_Catalog/).

국민들은 가공되지 않은 일차 데이터(raw data)를 보거나 다운로드 할 수 있을 뿐만 아니라, data.gov에서 제공하는 위젯이나 다른 툴을 이용하여 특정 데이터를 가공, 융합한 차트나 지도, 스냅샷을 만들 수도 있다(오원석, 2009).

3.2 영국

영국 정부는 공공부문의 정보공유 및 활용에 따른 가치창출을 위해 '정보의 힘(Power of Information, POI)' 보고서를 기반으로 2010년 1월 21일 공공정보 공개사이트(www.data.gov.uk)를 구축하였다.

영국 총리에 의해 구성된 공공부문투명성위원회(The Public Sector Transparency Board)가 제시한 14개 원칙들은 다음과 같이 요약할 수 있다. 1) 공공정보정책은 언제, 어떤 형태로든 대중과 기업의 필요에 따라 적용되어야 한다. 2) 공공정보는 재사용이 가능하며 기계가독형으로 생산되어야 한다. 3) 공공정보는 상업적 재사용을 포함해서 무료로 재사용이 가능한 오픈 라이선스로 공개돼야 한다. 4) 공공정보는 통일되고 접근이 용이한 형태의 온라인 창구를 통해 쉽게 이용할 수 있어야 한다. 5) 공공정보는 공개표준과 W3C의 권고안을 따른다. 6) 동일 주제를 다루는 서로 다른 부처의 공공정보는 표준 포맷으로 출판되어야 한다. 7) 정부 소유 웹사이트의 공공정보는 데이터와 서비스에 접근하여 재사용이 가능하도록 출판되어야 한다. 8) 공공정보의 공개는 시의적절해야 하며, 생산 직후 공개해야 한다. 9) 데이터를 빨리 개방해서 오픈 표준포맷으로 이용가능하다는 확신을 줄 수 있어야 한다. 10) 공공정보는 법적으로 자유롭게 이용가능해야 한다. 11) 공공정보는 특정 어플리케이션이나 등록절차 없이 이용할 수 있어야 한다. 12) 공공단체는 공공정보를 재사용하는 데에 적극적이어야 한다. 13) 공공단체는 데이터 자산목록을 관리하고 출판해야 한다. 14) 공공단체는 데이터셋에 대한 적절한 메타데이터를 생산해야 하며 이것은 통일된 온라인 접근점을 통해 이용할 수 있어야 한다(UK Cabinet Office, 2012).

미국/영국의 사례에서 보듯이 정부 분야의 공공데이터를 Linked Data화하는 작업이 주요하게 진행되고 있다.

4. 가이드라인 설계

국가나 지방자치단체, 공공기관이 보유한 공공정보의 개방과 재사용을 통해 사회경제적 가치를 창출하도록 적극적인 정보개방 의지가 필요하며, 공공정보 요청과 활용은 시민의 권리라는 것을 인지하고 공공정보를 보유, 관리, 제공하는 공공기관의 인식을 변화시키는 것이 중요하다(이정아, 2010).

LOD 구축을 위한 기술적인 개요는 그림 1과 같고 데이터, 시스템, 서비스 표준화로 나누어서 살펴보기로 한다.

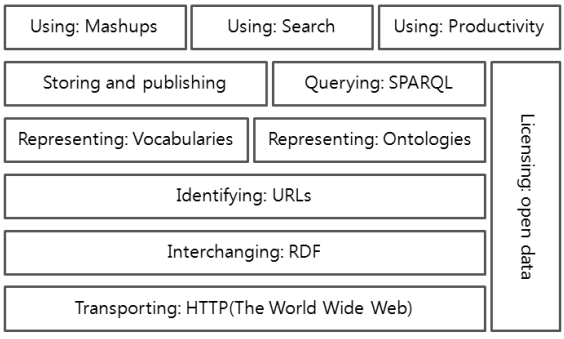

그림 1. LOD 구축과 활용에 관한 전략(Tim Davies. Elements of the Linked Open Data Puzzle 2011. Bauer 등. 'Linked Open Data: the Essentials' p.30에서 재인용)

4.1 데이터 표준화

이 연구는 기본적으로 국가나 공공기관이 보유하고 있는 정보자원의 공개를 통한 국가

경쟁력 제고를 위해 개방형 DBpedia를 구축하는 것이다. 따라서 데이터 표준화는 국가의 표준으로 사용하는 하나의 지침이 될 수 있다. 현재 우리 정부는 아담스라는 데이터베이스 표준화 지침을 모든 정부기관과 공공단체에 하나의 지침으로 제시하고 있다. 그에 따르면 우리나라 공공정보에 대한 데이터 표준화 원리를 다음과 같이 전제하고 있다(행정안전부, 2006).

- “행정 데이터 관리시스템(이하 관리시스템)”이라 함은 이 지침에 따라 산출되는 결과물을 저장·관리·확인하는 자동화된 도구를 말한다.
- “데이터요소”라 함은 논리적 데이터요소와 물리적 데이터요소로 구성된다.
- “행정표준용어사전”이라 함은 각급 행정기관에서 업무적으로 사용되고 있는 명사형 어휘에 대해 행정정보데이터베이스 구축·운영 시 사용하기 위하여 분류하고 정의한 어휘들의 집합을 말한다.

한편 LOD의 아키텍쳐는 일반적으로 두 가지 원칙을 따른다. 하나는 데이터 표현 및 접근에 관한 표준을 준수하는 것이다. 예를 들면, Open API를 이용하여 데이터를 수집하는(harvesting) 방식은 제공자가 정해 놓은 데이터만 전달된다. 이에 비해 Linked Data는 표준화된 데이터모델과 데이터 접근 메커니즘 및 스스로 데이터를 설명하는 형식으로 데이터구조에 대한 정보와 함께 데이터의 원소스의 수준을 완전하게 수요자 입장에서 배포한다. 두 번째는 공공정보를 개방한다는 것은 데이터의 상용화나 재가공에 아무런 제약이 없는 완전 공개된 데이터라는 점이다. Linked Data 아키텍처는 개방적이고 새로운 데이터를 실시간에 검색하게 한다. 이것은 신규데이터가 즉시 웹데이터로 활용하게 하는 장점을 갖는다. 공공정보에 오픈 라이선스를 도입한다면 공공정보의 자유로운 이용을 도모하여 공공정보의 자유이용허락이 확산되는 효과를 얻을 수 있고, 공공기관이 이용허락 조건을 공공정보에 스스로 명시함으로써 민간에게 공공정보를 적극적으로 활용할 수 있는 유인책으로 작용할 수 있다(최진원, 2011).

데이터 수집의 표준화는 이러한 점을 고려하여 DB공여자들은 다음 3단계 절차(CKAN, 2012)로써 데이터를 구조화하도록 한다.

- 레벨 1 : Basic level – Name, Title, URL, Author, Author email
- 레벨 2 : Minimal level – added tag(ex. topic), added resources(RDF ex link) SPARQL endpoint, new keys (ex. triples)
- 레벨 3 : Complete level – advanced info (ex. Version, Notes, Licence), new keys(namespaces etc), Resources(RDF Schema, Vocabulary Mappings)

4.2 시스템 표준화

LOD 데이터에 관련된 시스템은 크게 브라우저와 검색엔진으로 구분한다. 현재 Linked Data브라우저는 DBpedia에서 제공하는 DB간의 RDF 링크를 기점으로 제공된 원데이터 소스를 통해서 접근하는 형태를 취하고 있다. Linked Data 검색엔진은 해당 DB에서 제공하는 링크를 기점으로 Linked Data에 망라적으로 접근하여(크롤링) 수집된 데이터(하베스팅)를 하나의 결과대상 DB화하여 질의업무를 수행한다. 이 검색엔진들은 검색된 데이터에 대한 데이터구조를 설명하여야 하며 사용자에게 주제에 대한 링크 확장을 통해 풍부한 상호운용성을 제공하여야 하는 플랫폼을 구축하여야 한다. 이를 위해 제공되는 플랫폼은 대용량 트리플의 저장을 비롯하여 관리와 함께 질의처리를 담당하는 트리플 스토어 형태를 유지한다.

- Linked Data 플랫폼은 핵심적인 트리플 스토어를 기반으로 기존 레거시 데이터의 트리플화(RDF, 시맨틱 데이터)를 위한 트리플 변환기, 추론을 위한 추론기와 함께 Linked Data를 운영, 관리하기 위한 관리 환경을 구성한다.
- 기존의 오픈소스 웹서버를 포함하는 방식으로 웹서버 환경과 연동하는 모델이어야 한다.
- RDF 트리플 형태의 서비스 이외와 함께 RDFa 형태로 웹 사이트에 게시하도록 한다.
- 플랫폼은 SPARQL Endpoint를 기본으로 제공하여야 한다. 또한 SPARQL 질의 사용이 어렵거나, 해당 Linked Data 서비스의 모델을 인지하지 못하면 캐쉬 구조의 사용을 증진하기 위한 중간 매개체로 REST 기반의 API를 제공하도록 한다.

이러한 절차표준화는 Linked 브라우저와 검색엔진을 고려할 때 최적의 시스템 플랫폼이 갖추어야 할 기준이라 할 수 있다.

4.3 서비스 표준화

Linked Data 서비스는 다음의 4가지 원칙을 공통적으로 따른다. 첫째, Linked Data는 특정 개념을 URI로 명명한다. 둘째, 이름을 명명할 때는 인간들이 상대적으로 용이하게 식별할 수 있도록 HTTP URI를 사용한다. 셋째, URI는 표준 RDF 나 SPARQL을 사용해서 정보를 제공해야 한다. 넷째, RDF에 포함되어 있는 다른 정보자원에 접근할 수 있어야 한다. 이용자들이 더 많이 끊임없이 유용한 정보를 발견할 수 있도록 URI로 연결될 수 있는 링크를 제공해야 한다.

Linked Data의 서비스는 크게 네 가지의 유형으로 구분할 있다. 첫째, 일련의 캐털로깅 수준의 정보(목록정보로써 소재정보)만을 제공하는 것이다. 두 번째, RDF/XML과 같은 파일 형식으로 제공하는 것이다. 세 번째, 관계형 데이터베이스에서 Linked Data제공하는 방식으로 RDBMS 관점에서 Linked Data를 변환하여 배포하는 것이다. 넷째, RDF트리플(triple) 저장소로부터 Linked Data의 제공하는 방식이다. 모든 RDF 트리플들은 Linked Data 인터페이스까지도 제공하는 것이다.

5. 결론

이 연구는 공공정보로써 국가가 생산한 정보원들을 체계적으로 공개하여 민간과 정부가 재사용할 수 있는 방안가운데 하나인 절차표준화에 대한 연구이다. 이와 같은 공공정보개방을 위해서는 위키피디아와 연계된 디비피디아(DBpedia) 구축을 우리나라에서 구조화한 한국형 플랫폼과 이를 통한 서비스를 제공하는 절차표준화를 제안하였다. 제안은 크게 데이터에 관한 것과 시스템에 관한 것, 서비스에 관한 것으로 제안하였다. 실제로 이와 같은 표준제안이 학술적으로 가치를 갖기 위해서는 실제 국가 DBpedia 플랫폼을 설계하고 이를 연계하는 과정이 수용되는 실제 절차를 거쳐 그 수용여부를 검증할 필요가 있다. 따라서 차후 연구는 실제 우리나라 공공정보와 민간정보까지 확대한 국가지능형 데이터베이스 연계시스템에 대한 것이다. 이 때 추가로 부착할 절차는 '거버넌스'의 부분과 공공기관의 '권한과 책임'에 대한 절차도 함께 고려하여 개발할 것이다.

참고문헌

노영희. 2012. dCollection의 링크드 데이터 구축에 관한 연구. 한국도서관정보학회지: 43(2); 247-271

오원석. 2009. 소프트파워 창출을 위한 Linked

Data 기반 e-Government 추진전략. TopQuadrant Korea, 2009. 10.

이정아. 2010. 스마트 정부의 공공정보 개방과 이용활성화 전략. 한국정보문화진흥원 CIO Report Vol.28

최진원. 2011. 공공정보의 자율적 개방 확산을 위한 제도 도입 및 적용방안 연구. 한국정보문화진흥원 NIA VII-RER-11100

최진원. 2012. 공공정보 이용활성화를 위한 법제도적 과제에 대한 연구. 정보법학: 16(1); 237-266

한국문화정보센터. 2012. 지능형 웹기반 문화정보 활용 활성화 연구보고서 [cited 2012 Jul 25]. Available from: www.kcis.or.kr/info/down_pds.asp?no=1554

행정안전부. 2006. 행정정보 데이터베이스 표준화지침(안). 행정안전부

Bauer F, Kaltenbock M. Linked Open Data: the Essentials [cited 2012 Jul 20]. Available from: http://www.semantic-web.at/LOD-TheEssentials.pdf

CKAN. 2012. Guidelines for Collecting Metadata on Linked Datasets in CKAN [cited 2012 Jul 21]. Available from: http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/CKANmetainformation

Gill L, et al. 1993. Computerised linking of medical records: methodological guidelines. Journal of Epidemiology and Community Health 1993: 47(4); 316-319

UK Cabinet Office. 2012. Open data white paper: unleashing the potential [cited 2012 Jul 5]. Available from: http://www.cabinetoffice.gov.uk/resource-library/open-data-white-paper-unleashing-potential