

# 웹 기반 데이터 표준화 동향



김학래 한국과학기술정보연구원(KISTI) 재난정보서비스연구실 박사

## 1. 머리말

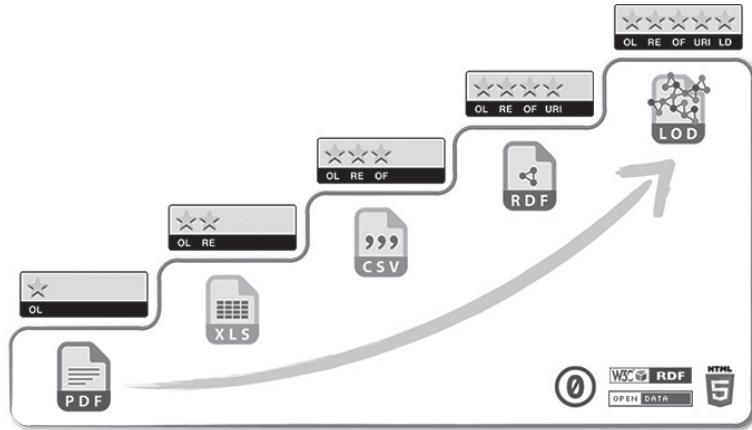
월드와이드웹(World Wide Web)은 문서의 연결을 통해 시작되었지만 대규모 데이터가 공유되는 공간으로 거듭나고 있다. 예컨대 위키피디아, 위키데이터는 데이터 전체를 다운로드 가능한 형식으로 제공하고 있고, data.gov 사이트(<https://www.data.gov/>)는 미국 정부의 오픈 데이터를 제공하는 데이터 포털로써 주목받고 있다. 웹은 거대한 데이터베이스로 진화하고 있고, 이를 위한 기술 및 표준에 대해 주목할 필요가 있다. 본고는 웹에서 데이터 발행 및 소비를 위해 월드와이드웹 컨소시엄에서 제정한 권고안을 소개한다. 특히 기계가 데이터를 판독할 수 있는 기능을 위한 표준을 자세히 살펴본다.

## 2. 판독이 가능한 데이터

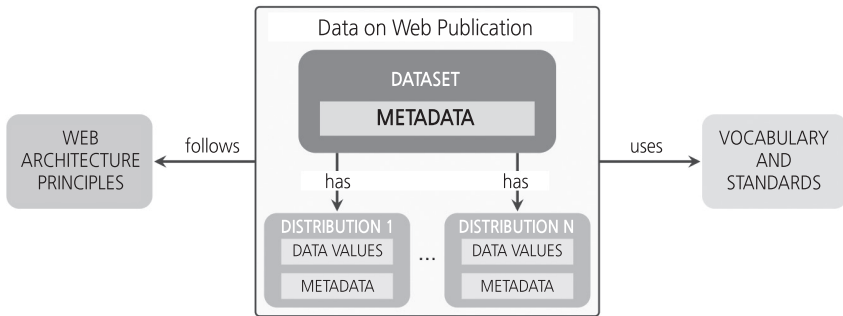
판독(read, 判讀)의 사전적 의미는 뜻을 헤아려 글을 읽는 것이다. 데이터 관점에서 보면, 파일이나 데이터베이스에 있는 데이터를 소프트웨어로 읽을 수 있는 것으로 해석할 수 있다. [그림 1]에서 보듯이, 기계가 판독하기 위한 수준은 여러 단계로 구분

할 수 있다.

먼저 1단계는 인쇄할 수 있거나 하드디스크에 저장할 수 있는 데이터로 오픈 라이선스가 적용된 데이터이다(OL, open license). 2단계는 특정 소프트웨어를 이용해 데이터 수집과 시각화가 가능한 것으로 다른 포맷으로 데이터 변환 및 발행 가능한 것이다(OF, open format). 3단계는 특정한 소프트웨어 기능에 한정되지 않고, 원하는 방법으로 데이터 작업이 가능한 것이다(RE, machine readable). 4단계는 로컬 시스템이나 웹에서 데이터에 링크나 북마크가 가능하고 데이터 일부를 재사용할 수 있는 것이다(URI, uniform resource identifiers). 마지막 5단계는 웹에 있는 다른 자원과 연결되고 데이터를 자동으로 탐색 및 발견할 수 있는 형식이다(LD, linked data). 데이터 형식으로 보면, 1단계는 PDF 형식, 2단계는 HWP, XLS와 같이 한글이나 마이크로소프트의 엑셀이 필요한 것이고, 3단계는 CSV, JSON, XML 과 같은 오픈 포맷, 4단계는 RDF 형식이 해당한다. 마지막으로 5단계는 링크드 데이터 기술을 적용하여 데이터를 공개하는 것을 의미하는 링크드 오픈 데이터(Linked Open Data)를 의미한다.



[그림 1] 5-star 오픈 데이터



[그림 2] 웹 아키텍처 기반 데이터 발행 및 배포

<표 1> 웹 데이터 관련 권고안 현황(2017년)

날짜	제목	상태
2017년 1월 31일	Data on the Web Best Practices	Recommendation
2017년 9월 7일	Time Ontology in OWL	Proposed Recommendation
2017년 9월 7일	Semantic Sensor Network Ontology	Proposed Recommendation

### 3. 표준화 동향

#### 3.1 웹 기반 데이터(Data on the Web)

웹에 발행하는 데이터는 데이터 집합과 메타데이터의 조합으로 구성되고, 다양한 공간에 물리적인

로 배포된다(그림 2 참조). 데이터 집합은 웹 아키텍처 원칙을 준수하고 데이터 게시와 사용을 위해 특정한 어휘와 표준을 사용할 수 있다. 웹 데이터는 배포된 버전에 대한 이해가 다를 수 있기 때문에 구조적 메타데이터, 설명적 메타데이터, 액세스와 같이 신뢰할 수 있고 재사용에 기여할 수 있는 데이터

<표 2> 웹 데이터 발행 및 배포를 위해 활용할 수 있는 어휘

Prefix	Namespace IRI	Description
dcat	http://www.w3.org/ns/dcat#	Data Catalog Vocabulary(DCAT)
dct	http://purl.org/dc/terms/	Dublin Core Metadata Initiative(DCMI) Metadata Terms
dqv	http://www.w3.org/ns/dqv#	DWBP Data Quality Vocabulary(DQV)
duv	http://www.w3.org/ns/duv#	DWBP Dataset Usage Vocabulary(DUV)
foaf	http://xmlns.com/foaf/0.1/	Friend of a Friend(FOAF) Vocabulary
oa	http://www.w3.org/ns/oa#	Web Annotation Ontology
owl	http://www.w3.org/2002/07/owl#	Web Ontology Language(OWL)
pav	http://pav-ontology.github.io/pav/	Provenance, Authoring and Versioning(PAV)
prov	http://www.w3.org/ns/prov#	Provenance Ontology(PROV)
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	Resource Description Framework(RDF)
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema vocabulary(RDFS)
skos	http://www.w3.org/2004/02/skos/core#	Simple Knowledge Organization System(SKOS)

집합 및 배포에 대한 정보를 제공해야 한다. 특히, 개별 데이터 집합은 기계가 처리할 수 있는 형식으로 게시 및 배포되어야 한다.

웹 데이터와 관련된 표준은 다양한 영역에서 지속적으로 논의되고 있다. 2017년 1월 발표된 Data on the Web Best Practices 권고안은 데이터 발행 및 배포에 대한 포괄적인 모범 사례를 포함하고 있다. 이 권고안은 웹에서 데이터의 게시 및 사용을 위한 모범 사례를 제공한다. 모범 사례는 메타데이터, 데이터 라이선스, 데이터 출처, 데이터 품질, 데이터 버전, 데이터 식별, 데이터 형식, 데이터 어휘, 데이터 접근, 데이터 보존, 피드백, 데이터 강화, 재발행 등 14개 주제를 포함한다. 이 권고안에 따르면 데이터 게시와 사용은 사람과 기계가 발견하고 이해할 수 있어야 되는데, <표 2>와 같은 표준 어휘를 이용해 데이터 품질 정보, 출처 정보, 라이선스 정보 및 사용 정보와 같은 다양한 정보를 표현할 수 있다. 예를 들어, Data Catalog Vocabulary(DCAT)은 기계가 판독할 수 있게 데이터 집합에 대한 메타데이터를 기술하기 위한 표준으로, 웹에 발행된 데이터 카탈로그 사이의 상호작용을 위한 RDF 어휘를 제공한다. 한편 DWBP Data Quality Vocabulary

(DQV)는 DCAT으로 기술된 데이터 집합에 대한 데이터 품질을 표현할 수 있는 어휘를 제공한다.

한편, 2017년 7월 OWL Time 온톨로지와 시맨틱 센서 네트워크 온톨로지가 제안되었다. 먼저 OWL-Time 온톨로지는 세상에 존재하는 모든 객체와 관련된 시간적 속성을 설명하거나 웹 페이지에 기술된 시간적 개념을 표현하기 위한 OWL-2 DL 온톨로지이다. 이 온톨로지는 지속 시간에 대한 정보 및 날짜, 시간 정보를 포함한 시간적 위치와 함께 시간과 간격 간의 위상 관계에 대한 사실을 표현하기 위한 어휘를 제공한다. 시맨틱 센서 네트워크 온톨로지(SSN, Semantic Sensor Network)는 센서에 대한 정보 및 물리적 세계에 대한 정보를 기술하기 위한 온톨로지이다. SSN 온톨로지는 사물인터넷, 스마트 시티 등 센서를 기반으로 데이터를 수집하고 처리하는 다양한 영역에서 활용될 수 있다.

대규모 데이터를 웹에 발행할 때 기계가 판독할 수 있는 형식으로 제공하는 것이 매우 중요하다. 특히, 각국 정부에서 보유하고 있던 데이터를 개방하는 과정에서 오픈 포맷의 중요성이 강조된다. 오픈 포맷은 모든 소프트웨어에서 자유롭게 활용할 수 있는 형식의 데이터를 말하는데 CSV, JSON, XML

<표 3> 링크드 데이터 관련 권고안 현황(2017년)

날짜	제목	상태
2017년 5월 2일	Linked Data Notifications	Recommendation
2017년 2월 23일	Web Annotation Data Model	Recommendation
2017년 2월 23일	Web Annotation Protocol	Recommendation
2017년 2월 23일	Web Annotation Vocabulary	Recommendation

이 대표적인 예이다. 이런 관점에서 테이블 형식의 데이터를 CSV로 제공하기 위해 네 가지 표준이 이미 제정되었다(2015. 9월).

- **Model for Tabular Data and Metadata on the Web:** 테이블 형식의 데이터와 메타 데이터에 대한 기본 데이터 모델
- **Metadata Vocabulary for Tabular Data:** 테이블 형식 데이터에 주석을 달아주는 메타 데이터에 대한 어휘 정의
- **Generating JSON from Tabular Data on the Web:** 테이블 형식의 데이터를 JSON으로 매핑할 때 적용할 절차와 규칙 정의
- **Generating RDF from Tabular Data on the Web:** 테이블 형식의 데이터를 RDF로 매핑할 때 적용될 절차와 규칙 정의

### 3.2 링크드 데이터(Linked Data)


링크드 데이터는 HTTP, RDF, URI 등 웹 표준 기술을 이용하여 데이터를 구조화하여 의미적으로 서로 연결하기 위한 기술이다. 링크드 데이터는 데이터의 공개, 연계, 공유를 위한 기본적 요구사항으로 데이터의 의미적 표현과 상호운용성을 제공하기 위한 핵심 기술이다. 링크드 데이터와 관련된 표준이 지속적으로 표준으로 제정되고 있으며, 2017년에 다음의 표준이 제정되었다.

Linked Data Notifications 권고안은 RDF 데이터의 송신 및 수신하는 역할을 하는 응용 프로그램 간에 메시지 교환을 정의한 프로토콜이다. 이 프로토콜은 서로 다른 기술 스택에서 실행되는 알림의 송신자, 수신자, 소비자가 원활하게 함께 동작할 수 있어 분산화된 상호작용이 가능하다. LDN은 알림을

보내고 사용하기 위해 링크드 데이터 플랫폼(LDP, Linked Data Platform)을 특수하게 사용할 수 있다. 그러나 LDP의 모든 기능이 아닌 일부 기능의 집합으로 구현이 가능하다.

어노테이션(annotation, 주석)은 일반적으로 자원 또는 자원 간의 연관에 대한 정보를 전달하는 데 사용된다. 예를 들어, 웹 페이지에 대한 태그나 이미지에 대한 주석이 대표적인 예이다. Web Annotation Data Model은 다양한 하드웨어 및 소프트웨어 플랫폼에서 주석을 공유하고 재사용할 수 있는 구조화 된 모델 및 형식을 설명한다. Web Annotation Protocol은 웹 아키텍처 및 REST 모범 사례에서 활용할 수 있는 어노테이션을 작성하고 관리하기 위한 전송 메커니즘을 설명한다. 이 권고안에서 정의한 주석은 웹 어노테이션 데이터 모델과 웹 어노테이션 어휘의 요구 사항을 따른다. 마지막으로 Web Annotation Vocabulary는 Web Annotation Data Model에서 사용하는 RDF 클래스, 술어 및 명명된 엔티티 집합을 지정한다. 또한, 어노테이션 모델에서 사용되는 다른 온톨로지의 권장 용어를 기술하고 링크된 데이터 컨텍스트에서 JSON-LD 컨텍스트 및 웹 어노테이션 JSON 직렬화를 위해 사용된다. 요약해 보면, 어노테이션과 관련된 표준은 웹에 있는 모든 자원에 대한 메타데이터를 추가할 수 있는 기능을 제공하고, 동시에 링크드 데이터 기술을 활용해 어노테이션 데이터를 공유할 수 있는 규격을 정의하고 있다.

#### 4. 맺음말

인공지능, 빅데이터에 대한 관심과 함께 데이터에 대한 중요성은 지속적으로 강조된다. 그러나 특정한 환경이 아닌 웹에서 데이터를 공유하고 사용하는 과정에서 표준에 대한 이해가 필요하다. 정부에서 다양한 공공 데이터를 개방하고 오픈 데이터가 보편화될수록 웹 아키텍처와 웹 표준을 바탕으로 데이터를 구축하고, 공유하는 것이 중요하다. 특히, 기계가 자동으로 데이터를 읽고 처리하기 위한 환경에서 웹 데이터를 위한 표준 및 링크드 데이터 기술의 중요성이 강조된다. 그동안 링크드 데이터 또는 시맨틱 웹 기술은 검색이나 추론과 같은 사용자 서비스를 차별화하기 위한 목적으로 활용되는 경향이 있었는데, 웹 아키텍처 관점에서 데이터의 구조화, 의미적 표현 그리고 데이터 연결을 위한 기본적인 수단으로 이해하는 것이 바람직하다. 

#### [참고문헌]

- [1] Steve Speicher; John Arwe; Ashok Malhotra. W3C. Linked Data Platform 1.0. 26 February 2015. W3C Recommendation. URL: <https://www.w3.org/TR/ldp/>
- [2] Robert Sanderson. W3C. Web Annotation Protocol. W3C Recommendation. URL: <http://www.w3.org/TR/annotation-protocol/>
- [3] Robert Sanderson; Paolo Ciccarese; Benjamin Young. W3C. Web Annotation Vocabulary. W3C Recommendation. URL: <http://www.w3.org/TR/annotation-vocab/>
- [4] Robert Sanderson; Paolo Ciccarese; Benjamin Young. W3C. Web Annotation Data Model. W3C Recommendation. URL: <https://www.w3.org/TR/annotation-model/>
- [5] Deirdre Lee; Bernadette Farias Loscio; Phil Archer. W3C. Data on the Web Best Practices Use Cases & Requirements. 24 February 2015. W3C Note. URL: <https://www.w3.org/TR/dwbp-ucr/>
- [6] Manu Sporny; Gregg Kellogg; Markus Lanthaler. W3C. JSON-LD 1.0. 16 January 2014. W3C Recommendation. URL: <https://www.w3.org/TR/json-ld/>
- [7] Fadi Maali; John Erickson. W3C. Data Catalog Vocabulary (DCAT). 16 January 2014. W3C Recommendation. URL: <https://www.w3.org/TR/vocab-dcat/>

- [8] Riccardo Albertoni; Antoine Isaac. W3C. Data on the Web Best Practices: Data Quality Vocabulary. 15 December 2016. W3C Note. URL: <https://www.w3.org/TR/vocab-dqv/>
- [9] Ian Jacobs; Norman Walsh. W3C. Architecture of the World Wide Web, Volume One. 15 December 2004. W3C Recommendation. URL: <https://www.w3.org/TR/webarch/>
- [10] Simon Cox; Chris Little. W3C. Time Ontology in OWL. 07 September 2017. W3C Proposed Recommendation. URL: <https://www.w3.org/TR/owl-time/>
- [11] Armin Haller; Krzysztof Janowicz; Simon Cox; Danh Le Phuoc; Kerry Taylor; Maxime Lefrançois. W3C. Semantic Sensor Network Ontology. 07 September 2017. W3C Proposed Recommendation. URL: <https://www.w3.org/TR/vocab-ssn/>

#### [주요 용어 풀이]

- 공공데이터(public information(or open data)): 데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보.
- 시맨틱 어노테이션(Semantic Annotation): 특정 개체를 지식 베이스의 온톨로지와 매핑하여 추가적인 시맨틱 정보를 생성하는 것.
- 온톨로지(ontology): 정보화 시스템 내 각 분야의 공유된 개념화(shared conceptualization)에 대한 정형화되고 명시적인 명세(formal and explicit specification).