데이터맵 어휘

박하람, 송채은 중앙대학교 문헌정보학과 석사과정
김학래 중앙대학교 문헌정보학과 교수

1. 머리말

데이터는 정부나 기업에서 중요한 자산으로 인식하고 있다. 데이터는 소유의 개념을 넘어 누구나 사용하고 배포하는 공유의 대상으로, 웹을 통해 누구나 접근하고 가치를 증대하는 협업의 대상으로 변화하고 있다. 최근 정부와 기업이 제공하는 데이터 포털은 특정한 소프트웨어에 제약 없이 웹 표준을 기반으로 데이터를 제공하고 있다. 그러나 데이터 서비스가 다양해지고, 데이터의 규모가 증가함에 따라 분산적으로 존재하는 데이터에 접근하고 탐색하는 것은 어렵게 된다. 특히, 서로 다른 데이터 서비스 사이에서 데이터의 탐색과 연계를 위해 데이터를 구조적이고 의미적으로 표현하고 상호운용성을 지원하는 것이 필요하다.

본고는 데이터 서비스에서 사용되는 메타데이터를 기계가 읽을 수 있게 표현하기 위한 데이터맵 어휘의 특징을 소개하고, 활용 방법에 대해 구체적으로 살펴본다.

2. 데이터맵 어휘

2.1 어휘 재사용

데이터를 표현하기 위한 표준은 지속적으로 개발되고 있고, W3C를 중심으로 이미 다양한 어휘가 권고안(recommendation)으로 제정되어 있다. 데이터맵 어휘는 어휘 재사용을 설계 원칙으로 정의하고 있기 때문에, 필수적인 클래스와 속성에 한정해서 새롭게 정의한다. 자원(Resource)에 대한 정보를 기술하기 위해 RDF(Resource Description Framework), DCMI(Dublin Core Metadata Initiative)를 사용하고, 데이터세트는 DCAT(Data Catalog Vocabulary)을 포괄적으로 적용할 수 있다. 예를 들어, 데이터맵 어휘는 데이터 목록과 관련된 어휘를 자체적으로 정의하지 않고, DCAT(dcat:Catalog)과 Schema.

org(schema:DataCatalog)의 어휘의 사용을 권장한다. 데이터맵의 모든 구성요소는 유일한 식별자를 부여할 수 있도록 URI(Uniform Resource Identifier) 체계를 적용하고, 데이터 사이의 연결을 위해 그래프(graph) 구조로 표현한다. 데이터맵 어휘에서 재사용하는 어휘는 <표 1>과 같다.

2.2 개념모델(Conceptual Model)

데이터맵은 온라인에서 데이터를 제공하는 서비스의 운영과 관리에 대한 메타데이터를 표현하는데 목적이 있다. 데이터를 서비스하는 주체는 보유·제공하는 데이터의 목록, 데이터 유형, 제공기관 등 데이터의 관리적 측면과 데이터세트의 조회, 다운로드, 활용 등 운영적 측면의 정보를 포함한다. 최근 데이터 서비스는 정부와 민간을 포함해 다양해지고 있고, 데이터 서비스 사이에서 데이터를 공유하기 위한 방안에 대해 본격적으로 논의하고 있다. 데이터맵 어휘는 개별 데이터 서비스에 대한 메타데이터를 표준적인 방법으로 기술하고, 동시에 서로 다른 데이터 서비스의 정보를 통합하는 데 활용할 수 있다. 개념적으로 보면, 데이터맵은 데이터 서비스와 데이터 서비스가 갖고 있는 정보의 관계로 표현할 수 있다. 즉, 데이터맵 의 개념 모델은 다음과 같다.

데이터맵 DM은 C, A, R의 관계로 정의한다. 이때, C는 데이터 서비스의 집합, A는 C가 보유한 메타데이터의 집합이다. 메타데이터의 집합은 기관, 데이터세트, 서비스 유형과 같은 속성을 정의한다. R은 C와 A의 이진 관계다. 하나의 데이터 서비스는 정의된 메타데이터를 통해 확장할 수 있다. 예를 들어, 분류체계나 데이터형식과 같은 항목은 데이터 서비스에서 제공하는 정보이지만, 외부의 데이터와 개념적으로 연결될 수 있는 정보를 포함한다. 이와 유사하게 개별 데이터맵은 이종의 데이터맵과 연계되거나, 상위 수준의 데이터맵과 통합될 수 있다. 통합된 데이터맵 DM_G는 개별 데이터맵의 속성을 집합시킨 정보로 표현한다.

$$DM_G = (DM_1, DM_2, \ldots. DM_n)$$

개별 데이터맵 $DM_1, DM_2, \cdots$는 서로 연결되어 메타 수준의 데이터맵 DM로 통합시킬 수 있다. 통합된 데이터맵은 주제별 데이터 목록으로 구성하거나, 기관 사이의 대규모 데이터 목록을 통합하는 목적으로 구성할 수 있다. 예를 들어, 국가 수준의 데이터맵 DM_G는 개별 기관이 관리하는 이종의 데이터맵을 연계해 통합할 수 있다.

<표 1> 데이터맵에 사용된 네임스페이스

접두사(Prefix)	네임스페이스(Namespace)
dcat	http://www.w3.org/ns/dcat#
dcterms	http://purl.org/dc/terms/
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
xsd	http://www.w3.org/2001/XMLSchema#
schema	http://schema.org

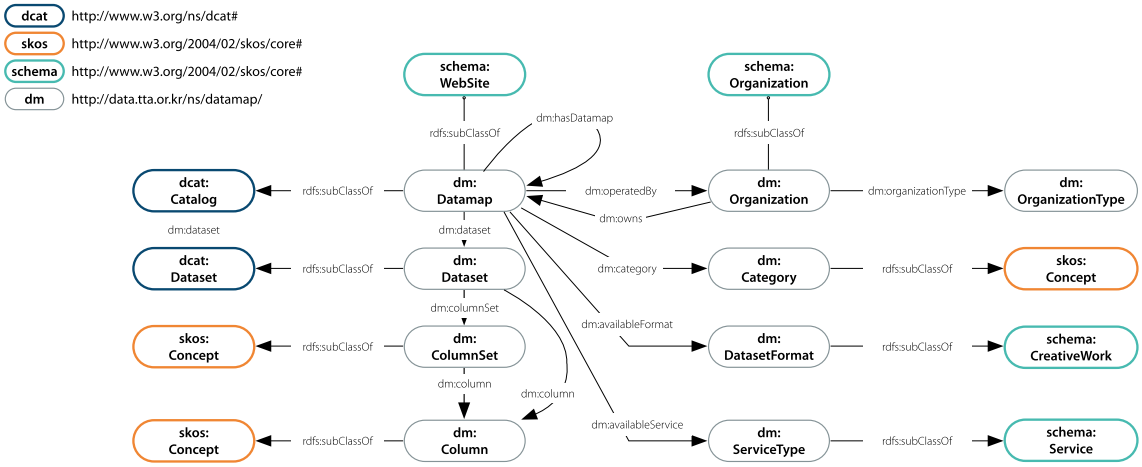
2.3 핵심 클래스와 속성

데이터맵 어휘는 dm:Datamap 클래스를 포함한 6가지 핵심 클래스로 이루어져 있다.

- **dm:Datamap**: 물리적인 데이터 서비스 안에 있는 모든 개체와 개체 사이의 메타데이터 관계를 표현한다.
- **dm:Category**: 특정 데이터 서비스가 제공하는 데이터의 주제 또는 분류를 나타내며, skos:Concept의 하위클래스다. 하나 이상의 주제가 있는 경우, skos:ConceptScheme 클래스를 이용하고, 개별 주제는 skos:Concept으로 표현한다.
- **dm:Organization**: dm:Datamap을 관리하거나, 구성하는 기관을 표현하는 클래스이다. dm:OrganizationType을 통해 기관의 유형을 구분할 수 있다.
- **dm:Dataset**: 데이터세트를 나타낸다.
- **dm:ColumnSet**: dm:Dataset이 가지고 있는 각각의 컬럼 모음이다. dm:Column이 가지고 있는 속성의 집합을 표현할 수 있다.
- **dm:Column**: dm:Dataset이 가지고 있는 개별 컬럼을 표현한다. 컬럼이 가지고 있는 속성(이름, 데이터타입, 설명 등)을 dcterms:title, dcterms:issued 등 기존의 어휘를 사용하여 관련 메타데이터를 표현할 수 있다.

데이터맵 클래스는 데이터 서비스를 제공하는 기관(dm:Organization), 제공하는 분류(dm:Category) 외에도 데이터세트의 유형(dm:DatasetFormat), 서비스의 유형(dm:ServiceType) 클래스와 연결되어 해당 정보를 표현할 수 있다. 데이터맵 어휘는 메타데이터 수준에서 데이터 목록을 기술하기 위한 속성을 정의하고 있다. dm:numberOfOrganization, dm:numberOfDataset, dm:numberOfCatalog, dm:numberOfColumn는 각각 데이터맵에 연계된 기관, 데이터세트, 데이터목록, 컬럼의 전체 개수를 표현한다.

데이터맵 어휘는 기존 어휘를 재사용할 수 있기 때문에, 어휘의 특징과 연계 방법에 대해 검토가 필요하다. 특히, 개별 데이터세트에 대한 정보는 DCAT 또는 Schema.org에서 정의한 어휘로 표현이 가능하다. 기존의 어휘로 데이터세트가 표현되었다면, dm:dataset 속성으로 dcat:Dataset 또는 schema:Dataset를 연결할 수 있다. 만약 데이터맵과 데이터세트를 새롭게 표현하는 경우, 데이터맵 어휘에서 정의한 dm:Dataset를 사용할 수 있다. dm:dataset 속성은 공역으로 dm:Dataset, dcat:Dataset,



[그림 1] 데이터맵 핵심 클래스와 속성

schema:Dataset을 갖는다.

3. 적용 사례

다음 예시는 데이터 포털과 같은 데이터 서비스에서 데이터맵을 적용하는 방법을 설명한다. 모든 예시는 Turtle 구문으로 표현한다. 데이터맵의 기본 정보는 <표 2>와 같이 표현할 수 있다. dm:catalog , dm:dataset 을 통해 데이터맵이 보유하고 있는 카탈로그와 데이터세트를 표현한다. 데이터맵의 주제분류와 기관은 각각 dm:category , dm:organization으로 표현한다. 이때 카테고리는 여러 유형을 가질 수 있다. 데이터맵이 보유하고 있는 데이터세트 수와 관련된 기관 수는 dm:numberOfDataset과 dm:numberOfOrganization으로 표현한다.

dm:Organization은 데이터 서비스를 관리하거나 데이터세트를 제공하는 주체를 표현하기 위한 클래스다. dm:owns는 해당 기관이 데이터맵을 관리하는 주체임을 표현하고, 복수의 데이터맵을 갖는 사실도 기술할 수 있다. dm:operatedBy 속성은 dm:owns 속성과 역 관계(inverseOf)로 데이터맵이 어떤 기관에서 관리되는지 표현할 수 있다. 또한, dm:Organization은 schema:Organization을 사용해서도 상세정보를 나타낼 수 있다. 기관에 대한 표현은 <표 3>과 같다.

데이터맵과 카달로그, 데이터세트의 연결은 <표 4>와 같이 기술할 수 있다. :datamap-001은 :catalog-001 , :catalog-002 , :catalog-0

<표 2> 데이터맵 클래스와 기본 속성

```
:datamap-001
    rdf:type dm:Datamap ;
    dm:catalog :catalog-001 ;
    dm:dataset :dataset-001, :dataset-002 ;
    dm:category :Education  ;
    dm:category :PublicAdministration  ;
    dm:organization :org-001 ;
    dm:numberOfDataset “2”^^xsd:NonNegativeInteger ;
    dm:numberOfOrganization “1”^^xsd:NonNegativeInteger .
```

<표 3> 기관의 표현

```
:org-001
    rdf:type dm:Organization ;
    dm:owns :datamap-001 ;
    dm:organizationType :OrganizationType/LocalGovernment  ;
    schema:email “hike.cau@gmail.com” ;
    schema:member “Haklae Kim” .

:datamap-001
    rdf:type dm:Datamap ;
    dm:operatedBy :org-001 .
```

<표 4> 데이터세트의 표현

```
:datamap-002
    rdf:type dm:Datamap ;
    dm:catalog :catalog-002, :catalog-003 ;
    dm:dataset :dataset-003, :dataset-004, :dataset-005 .

:catalog-002
    rdf:type dcat:Catalog ;
    rdfs:label “2st Imaginary Catalog”@en ;
    rdfs:label “두번째 가상의 카달로그”@ko ;
    dcat:dataset :dataset-003 .

:catalog-003
    rdf:type schema:DataCatalog ;
    rdfs:label “3nd Imagenary Catalog”@en ;
    rdfs:label “세번째 가상의 카달로그2”@ko ;
    schema:dataset :dataset-004 .

:dataset-003
    rdf:type dm:Dataset ;
    dm:numberOfDatasetColumn “10”^^xsd:NonNegativeInteger ;
    dm:numberOfDatasetRow “320”^^xsd:NonNegativeInteger .
```

<표 5> 데이터세트 컬럼의 표현

```
:dataset-003
    a dm:Dataset ;
    dm:columnSet [
        a dm:ColumnSet ;
        dm:column :column-001 ;
        dm:column :column-002 ;
        dm:column :column-003 .
    ] ;
    dm:column :column-001 ;
    dm:column :column-002 ;
    dm:column :column-003 .

:column-001
    a dm:Column ;
    rdfs:label "address"@en ;
    rdfs:label "주소"@koß ;
    dcterms:type <http://purl.org/dc/dcmitype/Text>  ;
```

<표 6> 데이터맵 사이의 연계

```
:datamap-002
    a dm:Datamap ;
    dm:hasDatamap :datamap-003, :datamap-004 ;
    dm:integratingStatus “true”^^xsd:boolean ;
    dm:numberOfCatalog “6”^^xsd:nonNegativeInteger ;
    dm:numberOfDataset “300”^^xsd:nonNegativeInteger ;
    dm:numberOfColumn “40000”^^xsd:nonNegativeInteger ;
    dm:numberOfColumnURI “28495”^^xsd:nonNegativeInteger .

:datamap-003
    a dm:Datamap ;
    dm:datamapOf :datamap-002 ;
    dm:url <http://3rd-datamap-example.org.html>  ;
    dm:numberOfCatalog “1”^^xsd:nonNegativeInteger ;
    dm:numberOfDataset “100”^^xsd:nonNegativeInteger ;
    dm:numberOfColumn “10000”^^xsd:nonNegativeInteger .

:datamap-004
    a dm:Datamap ;
    dm:datamapOf :datamap-003 ;
    dm:url <http://4th-datamap-example.org.html>  ;
    dm:numberOfCatalog “5”^^xsd:nonNegativeInteger ;
    dm:numberOfDataset “200”^^xsd:nonNegativeInteger ;
    dm:numberOfColumn “30000”^^xsd:nonNegativeInteger .
```

03을 목록으로 갖고 있고, 개별 목록은 각각 :dataset-001 , dataset-002 , dataset-003 의 데이터세트를 보유하고 있다. 데이터맵과 카탈로그는 dm:catalog 속성으로 표현하고, 카탈로그와 데이터세트의 관계는 dcat:dataset 속성으로 표현할 수 있다.

데이터세트는 하나의 컬럼 집합을 가지며, 컬럼 집합 dm:ColumnSet은 하나 이상의 컬럼을 가진다. 컬럼 집합은 rdf:Bag 의 하위 클래스이며, 다수의 컬럼을 가진 공백 노드로 표현된다. 컬럼은 컬럼 집합에서 dm:column으로 연결될 수 있고, 직접적으로 데이터세트에서 dm:column 으로 연결될 수 있다. <표 5>는 컬럼 정보를 표현하는 예시다.

데이터맵은 이종의 데이터맵과 통합할 수 있다. 2개 이상의 데이터맵이 결합한 데이터맵은 dm:integratingStatus를 통해 나타낼 수 있다. <표 6>에서 :datamap-002는 :datamap-003과 :datamap-004를 연계한 새로운 데이터맵이다. 통합된 데이터맵의 카달로그 수(dm:numberOfCatalog), 데이터세트 수(dm:numberOfDataset)와 컬럼 수(dm:numberOfColumn)는 :datamap-003과 :datamap-004의 중복을 포함한 합으로 표현된다. 중복 없는 컬럼 수는 dm:numberOfColumnURI로 표현할 수 있다.

X. 맺음말

본고는 데이터세트를 제공하는 서비스의 다양한 정보를 표준으로 표현하기 위한 어휘인 데이터맵을 소개했다. 데이터맵 어휘는 개별 데이터 서비스의 메타데이터와 더불어 복수의 데이터 서비스를 연계하고 통합하기 위한 방법을 제공한다. 데이터맵 어휘는 분산된 환경의 데이터 서비스에 대한 정보를 기계가 읽고 처리할 수 있는 형식으로 표현할 수 있다. 궁극적으로, 이종의 데이터 서비스에서 제공하는 데이터 현황을 쉽게 파악하고, 서로 다른 데이터 서비스 사이에서 데이터의 교환과 통합을 촉진시킬 수 있다.

주요 용어 풀이

- **SPARQL 엔드포인트**: SPARQL 1.1 권고안을 기반으로 데이터와 데이터 목록에 대한 연합 질의(federated query)를 지원하는 서비스
- **데이터맵**(Datamap): 데이터 서비스가 운영·관리하는 메타데이터와 이들 사이의 의미적 관계를 표현하기 위한 데이터 구조
- **데이터 서비스**(Data Service): 웹에서 데이터의 공개, 접근, 배포 등 데이터 활용을 지원하는 서비스 (예: 데이터포털, Github 등)
- **데이터 포털**(Data Portal): 사용자가 데이터세트에 접근할 수 있도록 지원하는 온라인 데이터 플랫폼
- **데이터 목록**(Data Catalog): 특정한 데이터 서비스가 보유한 데이터세트의 목록
- **데이터세트**(Dataset): 하나의 행에 하나의 관측치가 포함되도록 구성한 데이터의 집합
- **메타데이터**(Metadata): 데이터에 대한 데이터. 다른 데이터를 기술하기 위해 사용되는 데이터

참고문헌

[1] 김학래, 국가 데이터의 의미적 표현과 연계를 위한 데이터맵 지식 모델, 한국디지털콘텐츠학회 논문지, Vol. 22, No. 3, pp.491-499.

[2] 김학래, 국가데이터맵의 개념 및 모델, TTA Journal, Vol. 182, pp. 28-33, March. 2019.

[3] R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, P. Winstanley. Data Catalog Vocabulary (DCAT) - Version 2, 2020.

[4] ISA Programme of the European Commission, DCAT Application Profile for data portals in Europe Version 1.1 [Internet], 2021.

[5] DCMI Usage Board, DCMI Metadata Terms, Technical report, Dublin Core Metadata Initiative, Dec. 2006.

[6] R. V. Guha, D. Brickley, and S. MacBeth, Schema.org: Evolution of Structured Data on the Web: Big data makes common schemas even more necessary, ACM Queue, Vol. 13, No. 9, Nov-Dec, pp. 10-37, 2015.