JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 3, pp. 491-499, Mar. 2021

Check for updates

국가 데이터의 의미적 표현과 연계를 위한 데이터맵 지식 모델

김 학 래[1*]
[1*]중앙대학교 문헌정보학과 교수

A Knowledge Model of Data Map for Semantically Representing National Data

Hak-Lae Kim[1*]
[1*]Professor, Department of Library and Information Science, Chung-Ang University, Seoul 06757, Korea

[요 약]

인공지능, 빅데이터로 시작된 데이터 생태계의 급속한 변화는 전 세계 정부에 큰 도전으로 다가오고 있다. 대한민국 정부는 디지털 뉴딜, 데이터 댐을 통해 대규모 데이터 사업을 추진하고 있다. 그러나, 국가 차원에서 데이터의 관리 현황을 일관성 있게 제공하지 못하고 있다. 국내에서 데이터맵은 데이터의 출처 정보를 제공하는 서비스로 알려져 있지만, 데이터의 구조와 의미를 정의하지 않고 시각화에 집중된 경향이 있다. 본 연구는 대규모 데이터의 출처 정보를 의미적으로 표현하고 연계하기 위한 지식 모델을 제안한다. 데이터맵 지식 모델은 데이터 카탈로그 사이의 관계를 의미적으로 표현하기 위한 어휘이며, 서로 다른 데이터맵 사이의 연계·확장을 위한 개념 모형을 정의하고 있다. 데이터맵 어휘로 표현된 정보는 기계가 처리할 수 있기 때문에 이종의 데이터 포털 사이의 운영·관리 정보를 메타 수준에서 파악할 수 있는 방안이 될 수 있다.

[Abstract]

The rapid change in the data environment, which triggered by artificial intelligence and big data, is approaching a great challenge for governments around the world. The Korean Government is promoting large-scale data projects through Digital New Deal and Data Dam. At the national level, however, it has not been able to consistently provide the comprehensive status of data. In Korea, a data map is known as a service that provides provenance of individual data, but they tend to focus on visualisation without defining the structure and meaning of data. This study proposes a knowledge model for semantically expressing and linking provenance and relevant metadata of data from heterogeneous data services. The datamap knowledge model is a vocabulary for semantically expressing the relationship between data catalogs, and defines a conceptual model for linking and expanding heterogeneous data maps. Since the information expressed in the data map vocabulary can be processed by the machine, it can be a way to grasp information related to operation and management between heterogeneous data portals at the meta level.

색인어 : 데이터맵, 국가 데이터, 지식모델, 온톨로지, 데이터 연계

Key word : Data Map, National Data, Knowledge Model, Ontology, Data interlinking

http://dx.doi.org/10.9728/dcs.2021.22.3.491

Received 14 December 2020; **Revised** 05 February 2021
Accepted 05 February 2021

***Corresponding Author; Hak-Lae Kim**

Tel: +82-2-820-5561
E-mail: haklaekim@cau.ac.kr

Ⅰ. 서 론

전 세계적으로 데이터 경제의 중요성이 강조됨에 따라 국가별로 데이터 정책과 전략을 수립하고 있다 [1]. 국가별로 데이터 정책은 다를 수 있지만, 공공 부문의 데이터 (public sector data)는 데이터 정책의 핵심으로 인식되고 있다 [2]-[3]. 대한민국 정부는 공공기관이 보유하고 있는 공공정보를 '공공데이터'라는 이름으로 민간에 개방하고 있다. 공공데이터의 제공 및 이용활성화에 관한 법률[1]은 중앙부처, 지방자치단체, 산하기관의 공공데이터를 개방하는 근거 법률이다. 실제 이 법률을 근거로 공공기관은 공공데이터 제공 책임관과 실무자를 지정하고, 기관의 특성을 고려해 데이터 개방을 시행한다. 한편, 정부는 데이터 3법의 시행을 통해 데이터 환경의 근본적 변화를 유도하고 있다 [4]. 이런 측면에서, 공공데이터는 정부의 투명성 제고를 넘어 경제·산업적 가치에 대한 가치가 증가하고 있다.

반면, 수요자 측면에서 공공데이터의 가치는 여전히 높지 않다. 공공데이터의 규모는 지속적으로 증가하고 있지만, 사용자의 수요를 만족시키는 데이터의 발굴과 개방은 여전히 한계가 있다. 일반적으로 공공데이터의 한계는 수요 있는 데이터의 부족 [2], 낮은 수준의 품질이 원인이지만 [5]-[7], 정부가 보유한 데이터 현황을 파악하지 못하는 것은 다른 차원의 문제이다. 데이터의 보유와 소재 정보가 존재하지 않는 상황에서, 수요조사 중심의 접근은 단편적인 문제를 해결하는데 한정될 수 있다. 행정안전부에서 추진한 공공데이터 보유현황 전수조사[2]는 범정부 차원에서 데이터 소재정보를 파악하고 데이터 개방을 효과적으로 대응하기 위한 정책으로 평가된다. 그러나, 2018년에 시행한 전수조사는 전체 공공기관이 포함되지 않았고, 대규모 전수조사를 반복적으로 추진하는데 현실적 제약이 있다. 한편, 개방 데이터를 중심으로 진행된 조사 방식은 정부기관의 내부에 존재하는 데이터 현황을 파악하는데 한계가 있다. 이런 맥락에서 국가 데이터에 대한 개념적 정의가 필요하다. 공공데이터와 다르게, 국가 데이터는 공공기관이 보유·관리하는 데이터이고, 개방 여부에 관계없이 범정부의 모든 데이터를 대상으로 한다.

국가 데이터의 소재 정보는 데이터의 존재유무를 범정부 차원에서 인지하고, 운영과 관리 상황을 종합하여 데이터 관리체계를 수립하는데 필수적이다. 행정안전부는 공공데이터 보유현황 전수조사를 통해 확보한 결과를 범정부데이터플랫폼에 반영하고, '국가데이터맵[3]'으로 데이터 현황과 연관 데이터 정보를 제공하고 있다. 과학기술정보통신부는 10대 빅데이터 플랫폼에 포함된 데이터를 '통합데이터지도[4]'로 제공하고 있다 [8]. 그러나, 정부가 제공하는 데이터 출처 정보와 데이터맵은 데이터 거버넌스 관점보다 서비스 구현과 제공에 집중되어 있다. 실제, 행정안전부와 과학기술정보통신부에서 운영하는 데이터맵 서비스는 데이터 구조에 대한 정의 없이, 연관 데이터 정보를 시각적으로 보여준다. 따라서, 국가 차원의 데이터 보유 현황을 현재 구현된 데이터맵으로 표현하고 탐색하거나, 이종의 데이터맵 서비스의 데이터를 개방된 환경에서 연계·통합하는 것이 쉽지 않다.

본 연구는 국가 차원의 데이터 보유와 관리 현황을 의미적으로 표현하기 위한 어휘 체계를 제안한다. 데이터맵 지식모델은 데이터 관리주체가 독립적으로 관리하는 데이터의 소재와 연관정보를 표현하고, 이종의 데이터맵을 상호 연계하기 위한 의미 모델이다.

본 논문의 구성은 다음과 같다. 2장은 데이터맵과 관련된 표준, 서비스와 이론적 연구를 소개한다. 3장은 데이터맵의 필요성을 소개한다. 데이터맵과 관련 있는 법률과 서비스를 살펴보고, 개선 방안을 논의한다. 데이터 소재정보는 다양한 어휘로 표현될 수 있는데, W3C의 권고안인 데이터 카탈로그 어휘 (DCAT)를 소개하고, 데이터맵 지식모델과 차이를 기술한다. 4장은 데이터맵의 개념, 데이터 모델과 주요 어휘를 기술하고, 데이터맵의 적용 방안을 사례를 통해 소개한다. 5장은 연구내용을 요약하고, 향후 연구를 기술한다.

Ⅱ.관련 연구

데이터 목록 (data catalog)은 조직 또는 기관의 모든 데이터 자산을 탐색할 수 있는 메타데이터이며, 기계가 읽을 수 있는 (machine-readable) 형식으로 표현된다 [9]. 데이터의 급속한 증가로 메타데이터의 관리가 중요해짐에 따라 의미 기반 데이터 목록에 대한 필요성이 활발하게 논의하고 있다. W3C의 권고안 (recommendation)으로 제정된 DCAT (Data Catalog Vocabulary)은 웹에 게시된 데이터 목록 사이의 상호운용성을 제공하기 위해 설계된 어휘이다 [10]. 개별 데이터세트의 주요 정보가 DCAT으로 표현되면, 특정한 데이터 서비스에 제한되지 않고 분산되어 있는 데이터 목록을 검색할 수 있다 [11][12].

DCAT은 독립적으로 사용될 수 있고, 다른 어휘와 함께적용이 가능하다. Neumaier et al.은 영국과 미국의 데이터 포털에서 사용하고 있는 CKAN (Comprehensive Knowledge Archive Network) 메타데이터와 DCAT 어휘의 연계 규칙을 정의하고, DCAT 어휘의 실제 사용 사례를 검증하고 있다 [13]. DCAT-AP는 유럽의 공공 부분 데이터세트를 기술하기 위한 명세서로 DCAT을 기반으로 한다 [14]. 표 1과 같이, DCAT-AP는 데이터세트의 교환을 규격을 정의하고 있다. DCAT과 비교하면, DCAT-AP는 필수 어휘를 권고함으로써 데이터 연계를 위한 기준점을 제공한다. Jakub는 DCAP-AP를 체코의 국가 데이터 포털에 적용하기 위한 아키텍처와 데이터 표현 방안을 제안하고 있다 [15]. DCAP-AP를 적용하기 위한 그래픽 기반의 저작도구 [16], 국가별 확장 방안 [17], 데이터 포털의 적용 [18]-[22]에 대한 연구도 광범위하게 진행되고 있다. 한편, Schema.org도 DCAT과 유사한 성격의 클래스를 선언하고 있다. Schema.org는 검색엔진이 웹 사이트의 정보를 이해하는데 필요한 구조화된 데이터이다 [23].

1) https://bit.ly/2HnwoMK
2) https://bit.ly/2WglrQ1
3) https://www.data.go.kr/tcs/opd/ndm/view.do
4) https://www.bigdata-map.kr/

표 1. DCAT-AP의 주요 클래스
Table 1. Important Classes of DCAT-AP

Name	Description	Class URI	Mandatory
Agent	Objects associated with catalog and dataset. Agent contains individuals or organizations.	foaf:Agent	Yes
Catalogue	A catalogy or repository which describes datasets	dcat:Catalog	Yes
Dataset	A conceptual entity for representing published datasets	dcat:Dataset	Yes
Resource	All resources described by RDF	rdfs:Resource	Yes
Category	Subjects of datasets	skos:Concept	No
Category scheme	Controlled vocabularies about a category	skos:ConceptScheme	No
Distribution	A physical location of a dataset which is published in a certain format	dcat:Distribution	No
Licence document	A legal document about license	dct:LicenDocument	No

스키마 어휘는 보편적이고 범용적인 수준에서 어휘 체계를 구조적이고 의미적으로 정의하고 있다. 특히, 데이터세트와 데이터 목록은 각각 schema:Dataset, schema:DataCatalog 클래스로 정의하고 있다. 스키마 어휘의 데이터세트는 DCAT를 기본으로 하고 있지만, 기존 어휘의 재사용에 있어 다양한 논의가 진행되고 있고[5], 필요에 따라 매핑 규칙을 정의해 사용하기도 한다[6].

국내에서 데이터 목록에 대한 연구는 특정한 도메인에 적용한 방안이 소개되었지만 [24][25], 이론적 연구가 활발하지 않다. 반면, 공공 부문의 데이터 포털은 데이터 목록 서비스를 제공하고 있다. 예를 들어, 국가데이터맵은 700여개 공공기관의 개방 데이터의 소재 정보와 연관 데이터를 찾을 수 있도록 의미기반 검색을 지원하고 있다 [26]. 국가데이터맵에서 탐색한 데이터세트가 개방되어 있으면 바로 다운로드하고, 개방 예정인 데이터세트는 별도의 신청 절차를 통해 제공 받을 수 있다. 빅데이터플랫폼 통합 데이터지도는 금융, 환경, 문화, 교통, 헬스케어, 유통, 통신, 중소기업, 지역경제, 산림 분야의 빅데이터를 검색하고 소재정보를 제공하는 서비스이다 [8]. 10개 빅데이터 플랫폼이 개방하는 데이터 현황을 분야, 유형, 주제, 지역으로 구분해 시각화한 통계정보를 제공하고, 시맨틱 검색을 통해 데이터세트를 탐색할 수 있다. 데이터스토어[7]는 데이터를 온라인으로 판매하거나 구매할 수 있는 데이터 오픈 마켓으로, DCAT를 적용한 데이터 목록을 제공하고 있다.

대규모 데이터가 개방되고, 유통됨에 따라 데이터의 연계와 융합이 중요하게 인식되고 있다. 그러나, 독립적으로 운영하는 데이터 포털 사이에서 관련 있는 데이터를 검색하는 것은 쉽지 않다. 일부 데이터 포털이 데이터 목록 서비스를 제공하고 있지만, 데이터 세트의 의미와 구조를 표현하는데 한계가 있다. 대부분의 데이터 목록은 시각화 서비스에 집중된 경향이 있다. 본 논문에서 제안하는 데이터맵 지식 모델은 데이터의 소재 정보를 의미적으로 표현하기 위한 어휘를 정의하고, 메타 수준에서 이종의 데이터를 연계하는 방안이 될 수 있다.

5) https://schema.org/docs/data-and-datasets.html
6) https://github.com/w3c/dxwg/issues/251
7) https://www.datastore.or.kr/

III. 데이터맵의 필요성

3-1 데이터 관점의 데이터맵

행정안전부는 범정부데이터플랫폼에 데이터맵을 구축하고 있다. 공공기관의 데이터베이스 표준화지침에 따라, 개별 기관은 메타데이터 시스템을 구축하고 보유한 데이터베이스에 대한 43종의 메타데이터를 현행화하도록 권고하고 있다. 데이터맵은 범정부데이터플랫폼에 처음 적용되고 개별 플랫폼에 확산되고 있지만, 범정부 차원의 개념 모형을 정의하거나 표준화는 논의되지 않고 있다. 예를 들어, 과학기술정보통신부의 빅데이터플랫폼 통합 데이터지도 서비스는 행정안전부의 데이터맵과 유사한 성격이지만, 데이터맵 또는 데이터 지도 사이의 데이터 구조는 본격적으로 논의되지 않고 있다.

행정안전부의 데이터기반행정 활성화에 관한 법률(데이터기반행정활성화법[8])은 데이터맵의 법률적 근거가 될 수 있다. 이 법률은 데이터 기반의 과학적 행정 구현을 위한 추진체계와 기반 구축을 포함하고 있다. 특히, 데이터관리체계 마련을 위해 제4장 (데이터기반행정의 기반 구축) 제16조 (데이터관리체계의 구축)에 다음과 같은 세부 조항을 정의하고 있다.

① 공공기관의 장은 생성하거나 취득하여 관리하는 데이터에 대한 메타데이터(데이터의 체계적인 관리와 편리한 검색 및 활용을 위하여 데이터의 구조, 속성, 특성, 이력 등을 표현한 자료를 말한다. 이하 같다) 및 데이터관계도(데이터 간의 관계를 나타낸 그림을 말한다. 이하 같다)를 체계적으로 관리하여야 한다.
② 행정안전부장관은 데이터를 체계적으로 관리하기 위하여 공공기관의 메타데이터 및 데이터관계도를 통합·연계하여 관리할 수 있다. 이 경우 행정안전부장관은 기관별 메타데이터 및 데이터관계도를 종합하여 데이터관리체계를 구축·운영하여야 한다.

8) https://bit.ly/3lOy013

표 2. 국내 데이터 포털의 사례
Table 2. Examples of data portals in Korea

	Public data portal	Bigdata platform	Seoul Open data portal
Management	Ministry of the Interior and Safety	Ministry of Science and ICT	Seoul Metropolitan Government
Website	https://www.data.go.kr/	https://www.bigdata-map.kr/	https://data.seoul.go.kr/
Platform	Self-development	Self-development	Self-development
The number of files	37,483	9,638	6,607
The number of APIs	6,526	1,341	4,991
The number of classifications	18	10	12
The number of organisations	953	–	–

그러나 데이터기반행정활성화법에 정의한 '데이터 관계도'는 데이터의 관계를 시각화하는데 초점이 있다. 법률에 정의한 개별 조항은 데이터 관점에서 메타데이터, 데이터 관계도를 구현하는 구체적인 방안을 기술하지 않고 있다. 따라서, 범정부 차원의 데이터 요소, 구조적 특성, 연계를 위한 참조 모델(reference model) 등 데이터 관점에서 법률을 보완할 수 있는 방안과 공통모델의 표준 제정을 검토할 필요가 있다.

3-2 이종 데이터 목록의 연계

국가 수준에서 데이터 현황을 파악하는 것은 데이터 정책과 전략 수립에 필수적이다. 데이터 목록은 데이터의 출처정보를 요약할 수 있는 점에서 우선적으로 고려할 수 있다. 다만, 기관, 주제에 따라 독립적으로 운영되는 데이터 포털이나 서비스에 있는 데이터를 종합적으로 파악하기 위해 기술적 검토가 필요하다. 표 2에서 보듯이, 중앙부처, 지방자치단체는 개별적으로 데이터 포털을 운영하고 있다. 공공데이터포털은 행정안전부가 공공데이터의 개방을 목적으로 운영하고 있고, 과학기술정보통신부는 빅데이터의 구축과 유통을 위해 빅데이터 플랫폼을 구축하고 있다. 한편, 열린데이터광장은 서울특별시가 운영하는 공공데이터 서비스이다.

운영주체, 플랫폼이 다른 환경에서 다음과 같이 두 가지 질의에 대한 응답이 가능할까?

(1) 표 2의 데이터 포털에 공개된 파일 데이터의 전체 현황은?
(2) '교통' 관련 데이터 현황은?

질의 (1)은 개별 데이터 포털에서 파일 유형의 데이터 현황을 확인해야 하고, 질의 (2)는 데이터 분류의 유형이 '교통'으로 한정된 데이터를 파악해야 한다. 표 2에 있는 모든 데이터 포털이 서로 시스템 수준에서 연동되어 있다면, 두 가지 질의의 결과는 비교적 수월하게 얻을 수 있다. 그러나, 데이터 포털 사이의 시스템 연계가 현실적으로 쉽지 않다. 실제 중앙부처, 지방자치단체가 운영하는 데이터 포털의 전반적인 상황은 보고되지 않고 있다.

한편, 모든 데이터 포털이 DCAT을 지원하는 상황을 고려할 수 있다.

표 2의 데이터 포털이 DCAT으로 데이터 목록, 데이터세트를 제공한다고 가정해도, 분산적으로 존재하는 데이터 목록의 검색은 두 가지 조건을 만족시켜야 한다. 첫째, 데이터 서비스 제공자는 실시간 엔드포인트 (endpoint) 서비스를 지원해야 한다. 둘째, 엔드포인트 서비스는 SPARQL 1.1 권고안에 정의된 연합질의 (federated query)를 지원해야 한다 [27]. 이론적으로 보면, 개별 데이터 목록에 접근하고, 데이터세트의 메타데이터를 추출하고 요약할 수 있다. 반면, 현실적인 이슈도 고려해야 한다. 단일 데이터 목록에 연결되는 데이터세트의 규모가 커지거나, 이종의 데이터 목록에서 정의한 메타데이터 규격이 다를 수 있다. 대규모 데이터세트가 있는 데이터 포털의 엔드포인트 서비스는 실시간으로 데이터의 현황을 수집하는데 효과적이지 않을 수 있다. 즉, DCAT은 데이터세트와 데이터세트 사이의 관계를 의미적으로 정의할 수 있지만, 독립적으로 운영되는 데이터 서비스의 메타데이터를 집합적 (aggregated)으로 처리하는데 효율적이지 않다. 여기서, 집합적이란 서로 다른 데이터 목록의 값을 통합시키는 것을 의미한다. 예를 들어, 질의 (1)의 결과는 세 가지 데이터 포털의 값을 모두 합한 53,728이다. 한편, 세 가지 데이터 포털이 SPARQL 1.1 권고안을 지원하지 않는다면, 시스템의 기능과 관계없이 분산된 엔드포인트의 질의는 수행할 수 없다.

질의 (2)는 분류체계의 의미적 연계가 필수적이다. 표 2의 데이터 포털은 서로 다른 분류 체계를 정의하고 있기 때문에 의미적 연계가 필수적이다. 즉, 공공데이터포털 (교통물류), 빅데이터플랫폼 (교통), 열린데이터광장 (교통)의 분류정보는 의미적으로 동일하다는 정보와 식별자를 부여하고, 공통으로 적용해야 한다. 현재 데이터 포털은 이런 정보를 제공하고 있지 않기 때문에, 두 가지 질의를 처리하는 것이 쉽지 않다.

Ⅳ. 데이터맵의 개요

4-1 개념 모델

일반적으로 맵 (map)은 키와 값을 쌍으로 나열되는 자료 구조이고, 데이터 매핑 (mapping)은 서로 다른 데이터 모델 사이의 대응 관계를 갖는 데이터 요소를 말한다.

표 3. DCAT과 데이터맵의 비교
Table 3. Comparisons between DCAT and DataMap

	DCAT	DataMap
Objective	Metadata about datasets	Metadata about catalogs
Relationships	Relationships about datasets	Relationships about catalogs
Main class	dcat:Dataset, dcat:Catalog	dm:DataMap
Scope	Datasets on the Web	Data catalogs of a certain service or portal
Related standards	W3C recommendations and Schema.org	

본 연구에서 제안하는 데이터맵은 데이터 목록의 정보를 의미적으로 표현하기 위한 데이터 구조로 정의한다 [26]. 표 3에서 보듯이, 데이터맵은 DCAT과 표현 대상에 차이가 있다. DCAT은 데이터세트를 표현하기 위한 어휘이다. dcat:Dataset 클래스는 데이터세트에 대한 포괄적인 메타데이터를 기술하기 위한 속성 정보를 포함하고 있고, dcat:Catalog 클래스는 데이터 서비스 관점에서 다수의 데이터세트를 연계하는 기능을 제공한다. 데이터맵은 하나 이상의 데이터세트로 구성된 데이터 목록을 표현하기 위한 어휘이다. 따라서 데이터맵은 데이터세트를 표현하기 위한 별도의 어휘를 정의하지 않는다. 데이터맵의 개념 모델은 다음과 같다.

$$DM = (D, A, R) \quad (1)$$

데이터맵 DM은 D, A, R의 관계로 정의한다. 이때, D는 데이터세트의 집합, A는 D가 보유한 메타데이터 속성의 집합이다. 데이터세트의 속성은 개별 데이터세트에 따라 다를 수 있지만, 공통항목이 존재한다. 예컨대, '보유기관', '주제분류', '필드명'은 데이터세트를 이해하기 위한 주요 요소로 속성 집합에 포함될 수 있다. R은 D와 A의 이진 관계이다. 개별 데이터세트는 속성 정보를 통해 다른 데이터세트와 연결될 수 있다. 이와 유사하게 개별 데이터맵은 이종의 데이터맵과 연계되거나, 상위 수준의 데이터맵과 통합될 수 있다. 통합된 데이터맵은 개별 데이터맵의 속성을 집합시킨 정보로 표현한다.

$$DM_G = (DM_1, DM_2, ... DM_n) \quad (2)$$

개별 데이터맵 DM1, DM2은 서로 연결되어 메타 수준의 데이터맵 DMG로 통합시킬 수 있다. 통합된 데이터맵은 주제별 데이터 목록으로 구성하거나, 기관 사이의 대규모 데이터 목록을 통합하는 목적으로 구성할 수 있다. 예를 들어, 국가 수준의 데이터맵 DMG은 개별 기관이 관리하는 데이터 포털의 목록을 데이터맵으로 표현하고, 이종의 데이터맵을 연계해 통합적인 데이터맵으로 구성할 수 있다.

4-2 지식모델의 어휘

데이터맵 지식모델의 어휘는 필수적인 요소에 한정해 정의하고, 기존 어휘의 재사용을 원칙으로 한다. 자원에 대한 일반적 기술은 추가적인 어휘를 정의하지 않고, Resource Description Framework (RDF) [28], Dublin Core Metadata Initiative (DCMI) [29]를 사용한다. 데이터맵의 모든 구성요소는 유일한 식별자를 부여할 수 있도록 URI (Uniform Resource Identifier) 체계를 적용하고, 데이터 사이의 연결을 위해 그래프 (graph) 구조로 표현한다. 데이터 구성요소는 필수항목과 선택항목으로 구분할 수 있다. 필수항목은 데이터맵을 구성하는 핵심요소로써 보유기관명, 필드(컬럼)명이 포함될 수 있다. 특히, 필드명은 데이터의 연결 관계를 파악하는데 중요한 객체이며, 모든 항목이 정제되어 통제어휘 (controlled vocabularies)로 정의되어야 한다. 선택항목은 데이터 품질, 정부기능분류, 플랫폼 유형, 서비스 방식을 포함한다.

데이터맵의 핵심 클래스는 dm:DataMap이다. 이 클래스는 DCAT의 Catalog 클래스와 Schema의 DataCatalog 클래스와 동등한 관계로 정의한다 (owl:equivalentClass). 데이터맵은 하나 이상의 데이터 목록을 연계하거나 통합하기 위해 별도의 클래스를 정의한다. 즉, DCAT이나 Schema.org 어휘로 생성한 데이터 목록은 데이터맵을 통해 통합될 수 있다. 이때, 새롭게 생성되는 데이터맵은 데이터세트의 개별적 특성보다 데이터 목록 차원에서 집합적 관점으로 표현하고, 개별 데이터 목록이 갖고 있는 속성은 누적한 수치로 표현된다. 데이터세트의 표현은 별도의 클래스를 정의하지 않는다. 개별 데이터세트의 메타데이터는 DCAT이나 Schema.org의 클래스로 표현할 수 있고, 데이터 목록과 데이터세트의 관계는 dm:dataset 속성을 사용할 수 있다. 예를 들어, 데이터 목록과 데이터세트에 대한 정보가 DCAT과 Schema.org로 표현되었다면, 데이터맵은 별도의 연결을 고려하지 않는다. 반면, 데이터 목록과 데이터세트에 대한 의미적 표현이 존재하지 않는 경우, 데이터세트는 기존의 어휘를 사용해 표현하고, 데이터맵은 개별 데이터세트의 연결을 위해 dm:dataset 어휘를 사용한다 (그림 1 참고).

데이터맵은 메타 수준에서 데이터 목록을 기술하기 위한 속성을 정의하고 있다. dm:numberOfDataset, dm:numberOfOrg, dm:numberOfCatalog는 각각 데이터맵에 연계된 데이터세트, 기관, 데이터목록의 수치 정보를 표현하고 있으며, 하나 이상의 데이터 목록을 포함하면 해당 값은 모든 값을 합계해서 표현한다. dm:organizationType은 데이터맵을 운영·관리하는 기관을 표현하기 위한 속성이며, 해당 값은 정부 기관의 유형을 별도로 정의한다 (예: 중앙정부, 지방자치단체 등). dm:availableType 속성은 제공하는 데이터의 유형을 표시하기 위해 사용한다. dm:platform과 dm:dataQuality는 각각 데이터 목록의 플랫폼과 데이터 품질 정보를 표현하기 위한 속성이다.

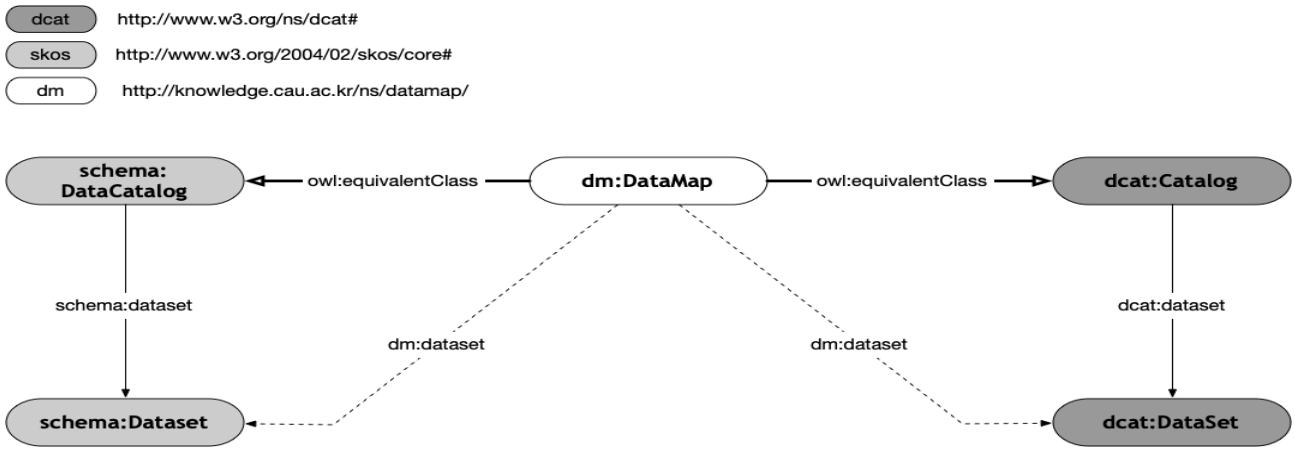

그림 1. 데이터맵 클래스
Fig. 1. The DataMap class

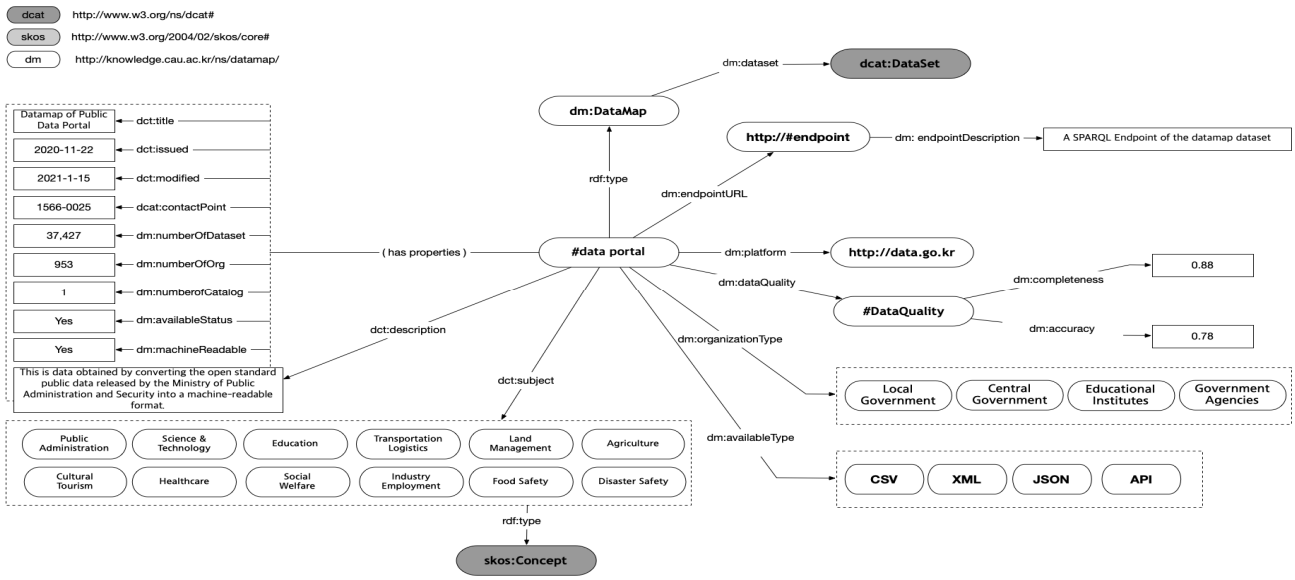

그림 2. 공공데이터포털 데이터목록을 위한 데이터맵
Fig. 2. The data map for describing data catalog of the public data portal

dm:columns 속성은 개별 데이터세트가 갖고 있는 컬럼명을 표현하는데 사용한다. 마지막으로, dct:title, dct:issued 등 다수의 기존의 어휘를 사용하여 관련된 메타데이터를 표현할 수 있다. 특히, dct:subject 속성은 데이터포털의 분류체계를 표현하기 위해 사용하며, 개별 분류체계는 Simple Knowledge Organization System (SKOS) [30]의 인스턴스로 정의할 수 있다.

4-3 구현 사례: 공공데이터포털

공공데이터포털에 대한 데이터맵 구현은 그림 2로 도식화하고 있다. '#data portal'은 데이터맵의 인스턴스이며, 공공데이터포털이 보유하고 있는 데이터세트의 현황과 운영·관리적인 정보를 포함하고 있다. 공공데이터포털의 상세 정보를 표현하기 위해 데이터세트의 규모 (dm:numberOfDataset), 개방기관 (dm:numberOfOrg), 서비스 상태 (dm:availableStatus), 기계판독 데이터의 제공여부 (dm:machineReadable) 속성을 적용하고 있다. 분류 체계 (dct:subject), 조직 유형 (dm:organizationType), 제공 데이터의 유형 (dm:availableType)은 해당 주제를 표현하기 위해 자원 값을 별도로 정의한다. 예를 들어, 주제 분류는 공공데이터포털에서 정의한 12개의 분류를 SKOS의 Concept 클래스의 인스턴스로 정의한다.

현재 공공데이터포털은 DCAT이나 Schema.org를 적용해서 데이터 목록을 표현하지 않고 있다. 따라서, 개별 데이터세트는 기존 어휘를 적용해 기술하고, 데이터맵 ('#data portal')은 해당 정보를 dm:dataset 속성으로 연결할 수 있다. dm:endpointURL과 dm:dataQuality 속성의 데이터 접근을 위한 URL과 데이터 품질을 표현한다. 데이터 품질 (dm:dataQuality)은 완전성 (dm:completeness)과 정확성 (dm:accuracy)을 기준으로 측정된 품질지수를 값을 사용할 수 있다. 그림 2의 데이터 품질 지수는 임의의 값을 표시하고 있다.

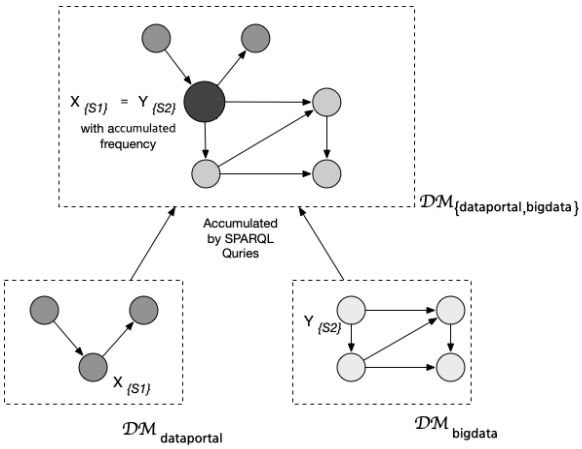

그림 3. 이종의 데이터맵의 연계와 통합
Fig. 3. Interlinking and integration between heterogenous data maps

4-4 데이터맵의 연계와 검색

데이터맵은 개별 데이터포털에 적용하거나 이종의 데이터맵을 연계해 새로운 데이터맵을 구성할 수 있다. 그림 3에서 보듯이, 데이터맵 DM {dataportal, bigdata}는 공공데이터포털과 빅데이터플랫폼 통합 데이터지도의 데이터맵 DMdataportal과 DMbigdata의 통합으로 생성된 메타 수준의 데이터맵이다. DMdataportal과 DMbigdata의 XS1과 YS2가 동일한 속성 정보를 갖고 있다면 DM {dataportal, bigdata}는 집합된 정보로 표현된다. 예를 들어, 데이터세트의 규모는 DMdataportal과 DMbigdata의 dm:numberOfDataset 속성의 값을 합계해서 집합적으로 표현할 수 있다. 표 4의 질의는 통합된 데이터맵이 갖고 있는 데이터세트의 규모를 검색하기 위한 SPARQL 질의문이다. 이 질의의 결과는 66,586개의 데이터세트로, 통합 데이터맵에 연계된 데이터세트의 전체 규모를 의미한다. 즉, 서로 다른 데이터맵에서 동일한 속성으로 표현한 값은 통합된 데이터맵에서 의미적으로 같은 정보로 해석한다.

반면, 분류체계와 같이 자원 유형 (resource)의 값을 갖고 있는 개체는 데이터맵 사이의 의미적 일치 관계를 먼저 정의하고 통합해야 한다. 먼저, 데이터 값의 URI를 식별할 수 있는 체계를 공통으로 적용해 유일성을 보장해야 한다. 공공데이터포털, 빅데이터플랫폼, 열린데이터광장의 분류체계는 동일하지 않다. 예를 들어, 교통과 관련된 주제는 '교통물류'와 '교통'을 사용하고 있기 때문에, 키워드가 아닌 URI 체계를 적용하여 유일성을 보장하는 것이 필요하다. 분류체계에 대한 URI 체계는 별도로 구성한다. '교통'의 URI는 'http://knowledge.cau.ac.kr/data/subject/transportation'로 정의하고, 통합 데이터맵에서 일치하는 키워드를 URI에 적용시킨다. 표 5는 통합 데이터맵에서 분류체계가 '교통'인 데이터세트의 규모를 확인하는 질의이다. 이 질의를 수행하면 개별 데이터맵에서 '교통'을 주제로 갖는 6,237개 데이터세트를 질의결과로 얻을 수 있다.

표 5. 이종의 데이터맵에서 특정 분류체계를 갖는 데이터세트 질의
Table 5. Query for datasets with a specific subject in heterogeneous datamaps

```
PREFIX rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:
<http://www.w3.org/2000/01/rdf-schema#>.
PREFIX schema: <http://schema.org/> .
PREFIX dm:
<http://knowledge.cau.ac.kr/ns/datamap/>.
PREFIX schema: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>.
PREFIX dct: <http://purl.org/dc/terms/>.
SELECT (SUM(xsd:integer(?dataset)) as ?datasets)
WHERE
{
   ?s dm:numberOfDataset ?dataset.
   ?s dct:subject
<http://knowledge.cau.ac.kr/data/subject/transp
ortation>.
        OPTIONAL {?s dm:avilableType
<http://knowledge.cau.ac.kr/data/format/json>.}

}
```

표 4. 이종의 데이터맵에 있는 전체 데이터세트 질의
Table 4. Query for total number of datasets in heterogeneous datamaps

```
PREFIX rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:
<http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>
PREFIX dm:
<http://knowledge.cau.ac.kr/ns/datamap/>.
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>.
PREFIX dct: <http://purl.org/dc/terms/>.
SELECT (SUM(xsd:integer(?org)) as ?orgs)
(SUM(xsd:integer(?dataset)) as ?datasets)
(SUM(?catalog) as ?catalogs)
WHERE
{
    ?s dm:numberOfOrg ?org .
    ?s dm:numberOfDataset ?dataset.
    OPTIONAL {?s dm:numberOfCatalog ?catalog.}
}
```

Ⅴ. 결 론

본 연구는 데이터 목록을 의미적으로 표현하기 위한 지식 모델을 제안하고 있다. 데이터 목록은 데이터세트에 대한 메타데이터를 기술하기 위해 사용되고 있다. 최근 공공과 민간에서 대

규모 데이터의 공유가 활발해짐에 따라 데이터의 출처정보를 제공하고, 효과적으로 탐색하기 위한 방안이 논의되고 있다. 특히, 데이터 목록은 서로 다른 데이터 포털 또는 서비스에 존재하는 데이터의 연계와 융합을 위한 주요한 수단으로 인식되고 있다.

본 연구에서 제안한 데이터맵 지식모델은 이종의 데이터 목록을 의미적 수준에서 연계하고, 메타데이터의 통합을 지원하는 방안을 제공한다. 데이터맵 어휘는 DCAT, Schema.org 어휘와 유사한 목적을 갖고 있지만, 하나 이상의 데이터 목록을 통합하고 집합적으로 표현할 수 있다. 한편, 데이터맵 지식모델은 개별 데이터세트를 직접 표현하기 위한 어휘를 정의하지 않고, 필요에 따라 연결할 수 있는 속성만을 정의함으로써 기존의 어휘의 재사용을 권장하는 설계원칙을 갖고 있다. 데이터맵 지식모델의 어휘는 DCAT을 포함한 W3C 권고안과 의미적 연결성을 갖고 있어, 웹 표준을 적용하여 데이터 목록을 구축할 수 있다.

본 연구는 이종의 데이터 목록을 연계·통합하기 위한 실험적 방안을 제안하고 있다. 국내에서 운영되는 대부분의 데이터 포털은 구조적·의미적으로 데이터세트와 데이터 목록을 표현하지 않고 있다. 데이터맵 지식모델은 개별 데이터 포털을 위한 데이터 목록으로 적용하거나, 다수의 데이터 목록을 통합하는 방식으로 적용할 수 있다. 더불어, 특정한 시스템에 종속적이지 않고 개방형 환경에 적용할 수 있기 때문에, 범정부 수준의 데이터 현황을 파악하기 위한 목적으로 데이터맵 지식모델의 적용을 검토할 수 있다. 개방된 데이터 포털이 아닌 정부 내부에서 활용하는 정보시스템에 적용해 공개여부에 관계없이 정부의 데이터 보유현황을 관리하기 위한 방안으로 검토할 수 있다.

향후 연구는 대규모 데이터 포털의 데이터세트를 DCAT이나 Schema.org로 표현하기 위한 방안, 데이터 목록 수준의 연계를 위한 기반 데이터의 범위와 설계 원칙이 필요하다. 개별 데이터 포털에서 정의한 메타데이터는 범정부 차원에서 통일하고, DCAT-AP와 같이 핵심 어휘를 정의하는 것을 검토해야 한다. 더불어 국가 데이터의 현황과 출처 정보를 효과적으로 파악하기 위한 방안과 데이터 거버넌스 체계에 대한 연구도 필요하다.

참고문헌

[1] Department for Digital, Culture, Media & Sport. Policy Paper: National Data Strategy [Internet]. Available: https://bit.ly/3kUHczy

[2] Deloitte. Open data: Driving growth, ingenuity and innovation, A Deloitte Analytics paper, 2012

[3] Organisation for Economic Co-operation and Development (OECD), “Government at a Glance 2019”, OECD Publishing, Paris, https://doi.org/10.1787/8ccf5c38-en, 2019.

[4] S.-A. Kim, “Meanings and Tasks of the Three Revised Bills which Ease Regulations on the Use of Personal Information,” Jouranl of Information and Security, vol. 20, no. 2, pp. 59–68, Jun. 2020.

[5] G. Kim, “An Evaluation of Public Data Opening Policy: Focused on Public Data Portal”, Public Policy Review, Vol. 31, No. 2, pp.57-82, 2017.

[6] H. L. Kim, “Quality Evaluation of the Open Standard Data”, Journal of The Korea Contents Association, Vol. 20, No. 9, pp. 439-447, 2020.

[7] H .Y. Kim, G. Y. Gim, “A Study on Public Data Quality Factors Affecting the Confidence of the Public Data Open Policy”, Journal of Information Technology Services (JITS), Vol. 14, No. 1, pp.53-68, 2015.

[8] S.T. Oh, “Major Projects of Data Dam and Future Directions,” ICT Issue Report, Vol.20. No. 03, pp. 1-12, Nov. 2020.

[9] Stephan, E. et al. “Semantic catalog of things, services, and data to support a wind data management facility.” Information Systems Frontiers 18 (2016): 679-691.

[10] R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, P. Winstanley. Data Catalog Vocabulary (DCAT) - Version 2 [Internet]. Available: https://www.w3.org/TR/vocab-dcat-2/

[11] Ponsoda, Elena Montiel and B. Villazón-Terrazas. “Impact of standards in European open data catalogues: a multilingual perspective of DCAT.” (2014).

[12] Heyvaert, Pieter et al. “Merging and Enriching DCAT Feeds to Improve Discoverability of Datasets.” ESWC (2015).

[13] Neumaier, Sebastian, Jürgen Umbrich, and Axel Polleres. "Challenges of mapping current CKAN metadata to DCAT." W3C Smart Descriptions & Smarter Vocabularies (SDSVoc) 2016.

[14] ISA Programme of the European Commission, DCAT Application Profile for data portals in Europe Version 1.1 [Internet], Available: https://bit.ly/2KdAS9M

[15] K. Jakub, “DCAT-AP representation of Czech National Open Data Catalog and its impact”, Journal of Web Semantics, Vol.55, pp.69-85, 2019.

[16] Rabissoni, Riccardo et al. “Metadata Editor: a web graphical tool for the DCAT-AP extensions RDF metadata generation.” (2019).

[17] Carlos Tejo Alonso, “DCAT-AP and its extensions: Context and evolution”, Technical report [Interner], Available: https://datos.gob.es/en/documentacion/dcat-ap-and-its-extensions-context-and-evolution

[18] Kirstein, Fabian et al. “Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP.” EGOV (2019).

[19] Cappello, Paolo, Marco Comerio, and Irene Celino. "BotDCAT-AP: An Extension of the DCAT Application Profile for Describing Datasets for Chatbot Systems." PROFILES@ ISWC. 2017.

[20] Perego, Andrea, et al. "GeoDCAT-AP: Representing geographic metadata by using the “DCAT application profile for data portals in Europe”." Joint UNECE/UNGGIM Europe Workshop on Integrating Geospatial and Statistical Standards, Stockholm, Sweden. 2017.

[21] Haller, Stephan, Beat Estermann, and Angelina Dungga Winterleitner. "Study in View of the Further Development of DCAT-AP CH." (2018).

[22] Dekkers, Makx, et al. "StatDCAT-AP, A Common Layer for the Exchange of Statistical Metadata in Open Data Portals." SemStats@ ISWC. 2016.

[23] 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.

[24] SHIN, Doh Kyoum, LEE, Sang Hwa, KANG, Junghyun, PARK, Eun Mi. (2019). Data Catalogue Standards Based on DCAT for Transportation Data: DCAT-Trans. Journal of Korean Society of Transportation, 37(5), 430-444.

[25] Jin Ho Park. “Designing Dataset Management and Service System for Digital Libraries Using DCAT”, JOURNAL OF THE KOREAN SOCIETY FOR LIBRARY AND INFORMATION SCIENCE, 53(2), 247-266, 2019.

[26] H. L. Kim, “The Concept and Model of National Data Map”, TTA Journal, Vol. 182, pp. 28-33, March. 2019.

[27] S. Harris, A. Seaborne. SPARQL 1.1 Query Language [Internet]. Available: https://www.w3.org/TR/sparql11-query/

[28] D. Brickley, R. V. Guha, “RDF Vocabulary Description Language 1.0: RDF Schema,” Technical report, W3C, http://www.w3.org/TR/2004/REC-rdf-schema-20040210/ 2004.

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

[30] Miles, A. and S. Bechhofer, SKOS Simple Knowledge Organization System Reference [Internet], Available: https://www.w3.org/TR/skos-reference/

김학래(Haklae Kim)

2010년 : 아일랜드 국립대학교 (공학박사)

1997년～2000년: Digital Enterprise Research Institute, Ireland

2009년～2016년: 삼성전자

2017년～2019년: 한국과학기술정보연구원

2019년～현 재: 중앙대학교 문헌정보학과 교수

※관심분야 : 지식그래프, 인공지능, 데이터 사이언스 등