JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 11, pp. 2249-2258, Nov. 2022

Check for updates

DCAT-AP-KR: 국내 데이터 포털의 상호운용을 위한 애플리케이션 프로파일

박 하 람[1] · 김 학 래[2*]
[1]중앙대학교 문헌정보학과 정보학 석사과정
[2*]중앙대학교 문헌정보학과 교수

DCAT-AP-KR: Application Profile for Interoperability of Data Portals in Korea

Haram Park[1] · Haklae Kim[2*]

[1]Master's Course, Department of Library and Information Science, Chung-Ang University, Seoul 06974, Korea
[2*]Professor, Department of Library and Information Science, Chung-Ang University, Seoul 06974, Korea

[요 약]

공공데이터는 데이터 경제를 확산하는데 매우 중요하다. 공공데이터는 데이터 제공자와 사용자의 수요와 공급의 지속적인 피드백을 통해 발전할 수 있다. 데이터 사용자는 공공데이터를 자유롭게 활용하고 서로 다른 데이터를 연계·융합해 새로운 가치를 만들 수 있다. 데이터 공급자인 공공기관은 사용자가 원하는 데이터를 개방하고, 더불어 서로 다른 데이터포털 사이의 연계를 위한 방법을 지원해야 한다. 그러나, 공공기관에서 운영하는 데이터 포털에서 사용하는 메타데이터에 대한 표준이나 가이드라인이 존재하지 않는다. 본 논문은 국내 데이터 포털에 공통으로 적용할 수 있는 메타데이터 요소 집합인 DCAT-AP-KR을 제안한다. 공공기관이 운영하는 데이터 포털에서 메타데이터를 추출하고 의미 분석을 통해 공통 메타데이터를 선정한다. 공통 메타데이터는 DCAT-AP 어휘의 설계 원칙을 적용하여 신규 어휘로 추가한다. DCAT-AP-KR은 데이터 포털에서 사용하는 메타데이터의 항목과 값을 일관되게 표현하는 어휘 명세로, 서로 다른 포털 사이의 의미적 연계를 가능하게 하고 동시에 분산된 데이터 포털에 있는 데이터세트의 탐색을 지원할 수 있다.

[Abstract]

Public data is increasingly critical for fostering data-driven economy. It can be improved through continuous feedback of supply and demand from data providers and consumers. Data consumers are able to freely use public data and create new values by interlinking and consolidating among heterogeneous data from different data sources. Public institutions, which are data providers, should offer a set of datasets that consumers need, and provide methods for identifying datasets among data portals, which aim to create a new dataset by interlinking them. However, there are no standards or guidelines for metadata used in data portals operated by public institutions. This paper proposes DCAT-AP-KR, a set of metadata elements that can be commonly applied to data portals in Korea. It extracts metadata from data portals operated by public institutions and defines a set of common metadata through semantic analysis. The common metadata is added as a new vocabulary by applying the design principles of the DCAT-AP vocabulary. DCAT-AP-KR is a vocabulary specification that consistently represent the common metadata elements and values by applying the design principles of the DCAT-AP vocabulary.

색인어 : 데이터세트, 데이터 포털, DCAT, DCAT-AP, 상호운용성
Keyword : Dataset, Data Portal, DCAT, DCAT-AP, Interoperability

http://dx.doi.org/10.9728/dcs.2022.23.11.2249

Received 17 October 2022; **Revised** 26 October 2022
Accepted 31 October 2022

***Corresponding Author; Haklae Kim**

Tel: +82-2-2275-4435
E-mail: haklaekim@cau.ac.kr

Ⅰ. 서 론

제3차 공공데이터 제공 및 이용활성화 기본계획[1]에 따르면, 공공데이터의 개방 전략은 공공기관 중심에서 국민이 자기 주도적으로 데이터를 활용하는 방향으로 전환하고 있다. 공공데이터의 전면 개방은 인공지능, 빅데이터 기술과 결합되어 데이터의 융합과 연계를 활발하게 만드는 촉매제가 될 수 있다.

공공데이터는 공공기관이 보유·관리하는 데이터이기 때문에 주제 범위가 매우 광범위하다. 기관에 따라 정책이 다르지만, 공공기관은 자체적으로 데이터 포털을 통해 데이터를 제공하고 있다. 그러나, 데이터 사용자가 서로 다른 데이터 포털에 있는 데이터를 연계하거나 통합하기 위한 방안은 본격적으로 논의되고 있지 않다. 공공기관에서 운영하는 데이터 포털은 메타데이터 항목명, 항목명의 값이 서로 다르고, 국가 차원에서 이를 규정한 방안이 존재하지 않는다. 행정안전부의 공통표준용어는 공공데이터에서 사용하는 컬럼명을 분석하여 범정부 표준용어로 제공하고 있다. 현재 535개 표준용어가 정의되어 있다[2]. 그러나, 개방된 공공데이터에서 표준용어의 활용이 미흡하고, 기존 컬럼과 매핑 체계를 제공하지 않기 때문에 데이터의 연계를 위해 사용하는데 제약이 있다. 한편, 공공데이터포털은 데이터세트의 메타데이터를 DCAT(Data Catalog Vocabulary)[3], Schema.org[4] 어휘로 제공하고 있다. DCAT은 분산되어 있는 데이터세트의 의미적 연계를 위해 제정된 W3C(World Wide Web Consortium)의 권고안(Recommendation)으로, 세계 각국의 정부와 민간 기관의 데이터 포털에서 광범위하게 활용되고 있다. 다만, 표준의 단편적 적용은 데이터 연계를 해결하는 데 효과적이지 않다. 예컨대, DCAT은 데이터세트와 관련 있는 다양한 정보 객체의 의미와 관계를 기술할 수 있지만, 공공데이터 포털은 대부분의 정보를 문자열(string)로 표현하고 있다. 즉, DCAT을 통해 표현된 정보는 기존의 데이터세트에 대한 메타데이터와 차별화되지 않고 형식과 구조만 변경했기 때문에, 사용자는 서로 다른 데이터 포털에서 데이터세트를 탐색하고, 관계성을 찾는데 활용하기 어렵다.

공공기관이 운영하는 데이터 포털은 서로 다른 메타데이터 체계를 갖고 있기 때문에, DCAT의 적용은 데이터 포털의 상호운용을 확보하는 방안을 함께 검토하는 것이 바람직하다. 이런 목적에서 유럽연합 집행위원회(European Commission)는 DCAT의 애플리케이션 프로파일(Application Profile: AP)을 제정해 데이터 포털에 적용하고 있다. 애플리케이션 프로파일은 특별한 애플리케이션 시스템을 위해 정의된 가이드라인 정책, 메타데이터 요소 집합으로 구성되며, DCAT-AP는 데이터세트에 대한 메타데이터 어휘를 구체적으로 정의하고 있다.

본 논문은 국내의 데이터 포털에 공통으로 적용할 수 있는 메타데이터 어휘를 검토하고, 신규 어휘를 적용한 DCAT-AP-KR을 제안한다. 공공기관이 운영하는 109개 데이터 포털에서 메타데이터 항목을 추출하고, 항목명의 의미를 분석하여 공통으로 사용하는 메타데이터를 선정한다. 공통 메타데이터는 DCAT-AP의 설계원칙에 따라 신규 어휘를 추가한 DCAT-AP-KR로 확장한다.

논문의 구성은 다음과 같다. 2장은 DCAP-AP 등 관련 연구를 소개한다. 3장은 공공기관이 운영하는 데이터 포털의 현황을 분석하고, 메타데이터 선정 방법을 기술한다. 4장은 공통 메타데이터의 추출과 선정 과정을 기술한다. 5장은 DCAT-AP 어휘를 기준으로 공통 메타데이터를 추가한 DCAT-AP-KR을 소개한다. 6장은 연구를 요약하고 향후 연구 방향에 대해 기술한다.

Ⅱ. 선행연구

DCAT(Data Catalog Vocabulary)은 분산된 웹 환경에서 공개되는 데이터 카탈로그를 기술하기 위한 RDF 어휘로, 데이터세트와 데이터 서비스를 기술하는 표준화된 모델을 제공한다[3]. 정부 데이터 카탈로그의 상호운용 맥락에서 시작된 DCAT은 2014년부터 W3C 권고안으로 채택되었고, 현재 DCAT 버전 3[5]이 제공되고 있다. DCAT이 적용된 데이터세트와 데이터 서비스는 기계가 읽고 이해할 수 있는(machine-actionable) 형식의 메타데이터로 기술된다. 이는 검색 엔진에서 데이터세트와 데이터 서비스의 탐색성을 높이고, 카탈로그 사이의 연합 검색(federated search)을 지원한다[5].

DCAT-AP(DCAT Application Profile for data portals in Europe)는 유럽의 데이터 포털에 DCAT을 적용하기 위한 구체적인 프로파일 모델을 제공한다[7]. 프로파일 모델은 유럽 데이터 포털 사이의 데이터세트 기술 교환을 용이하게 하기 위해 데이터세트의 메타데이터 레코드(metadata record)를 상세히 정의한다[6]. 특히, 프로파일 모델은 적용을 위한 의무(mandatory) 요소, 권장(recommended) 요소와 선택(optional) 요소를 지정하고, 기존 통제 어휘(예: EU Vocabularies)의 재사용을 권장한다[7]. 또한, DCAT-AP는 국가에 따라 어휘를 선별적으로 적용하거나 특정한 데이터세트를 표현하기 위해 확장하는 방식(extension)을 제공한다. 데이터세트 유형에 따른 익스텐션은 통계 데이터세트와 데이터 포털의 상호운용성을 강화하기 위한 StatDCAT-AP[8]와 공간 데이터세트 기술에 용이한 GeoDCAT-AP[9]가 있다. 국가에 따른 익스텐션은 국내 적용 요구에 맞추어 DCAT-AP 프로파일 모델을 확장한다. 국가에 따라 DCAT-AP의 의무, 권장, 선택 요소로 정의된 속성의 우선순위나 출현 횟수, 통제 어휘 등이 변동될 수 있다[10]. 대표적으로 독일(DCAT-AP.de)[11], 벨기에(DCAT-BE)[12], 네덜란드(DCAT-AP-DONL)[13], 스위스(DCAT-AP CH)[14], 이탈리아(DCAT-AP_IT)[15] 등이 DCAT-AP 익스텐션을 데이터 포털에 적용하고 있다.

유로피안 데이터 포털(European Data Portal)은 DCAT-AP를 사용하여 유럽의 공공데이터를 통합 검색할 수

있는 대표적인 예다. 현재 유로피안 데이터 포털은 36개국의 카탈로그 173개와 데이터세트 1,511,266개(2022년 10월 13일 기준)를 연합 질의할 수 있다[16]. 유럽의 각 국가는 데이터 포털의 공통 메타데이터 표준으로 DCAT-AP 또는 익스텐션을 사용하고, 메타데이터는 기계가 읽고 이해할 수 있는 형식으로 표현된다. 유로피안 데이터 포털은 각 국가의 데이터 포털로부터 메타데이터를 자동 이관 받고[17], 메타데이터 품질 측정을 자동화할 수 있다[18]. 결과적으로 RDF 기반의 메타데이터 표준 적용은 공공데이터의 탐색과 접근, 상호운용을 향상시키고, 데이터 포털의 통합에 드는 운영 비용과 시간을 감소시킬 수 있다[19].

국내 데이터 포털의 연구는 서로 다른 데이터 포털을 연계하여 데이터를 통합하려는 요구가 지속적으로 증가하고 있다[20]-[21]. 그러나, 데이터세트를 기술하는 가이드라인이나 표준이 부재하기 때문에, 데이터의 연계와 통합은 현실적 한계가 있다[22]. 데이터세트 수준의 연계와 상호운용을 위해 어휘 중심의 표준이 적용되어야 하지만, 국내는 데이터세트의 관리와 상호운용을 위해 DCAT을 적용하려는 시도가 초기 단계에 있다[23]-[28]. [29]에서 언급하듯이, DCAT의 적용은 국내 데이터 포털이 제공하는 메타데이터를 먼저 검토하고, 국내 포털에 맞게 확장하는 것이 논의되어야 한다. 본 논문은 국내 데이터포털에서 사용하는 공통 메타데이터 항목을 추출하고, 이를 반영하여 국내 데이터 포털을 위한 표준 어휘 DCAT-AP-KR을 제안한다.

Ⅲ. 연구 방법

3-1 조사 대상의 선정

본 연구를 위한 데이터 포털의 선정은 그림 1과 같은 절차를 통해 수행된다. 첫째, 공공기관이 운영·관리하는 데이터 포털의 현황을 파악하고, 수집 범위를 선정한다. 2022년 2월 1일 기준으로 공공기관은 109개의 데이터 포털을 운영하고 있다. 둘째, 본 연구는 데이터 포털에서 제공하는 기술적인 메타데이터 (descriptive metadata)를 대상으로 하기 때문에, 특정한 주제 분야의 데이터 포털은 수집에서 제외한다. 예를 들어, 통계와 공간정보와 관련된 데이터세트는 주제 영역의 고유한 메타데이터를 다수 포함하고 있기 때문에, 이와 관련된 41개 데이터 포털은 제외한다. 셋째, 데이터세트에 대한 메타데이터가 제공되지 않는 23개 데이터 포털을 제외한다. 국민건강영양조사포털은 데이터를 개방하고 있지만, 데이터세트에 대한 메타데이터를 제공하지 않는다. 울산광역시 데이터포털은 공공데이터포털에 데이터세트를 제공하고 해당 데이터세트의 URL을 제공하기 때문에 조사대상에서 제외한다. 표 1은 조사대상으로 선정된 공공기관과 운영하는 데이터 포털에 대해 요약하고 있다.

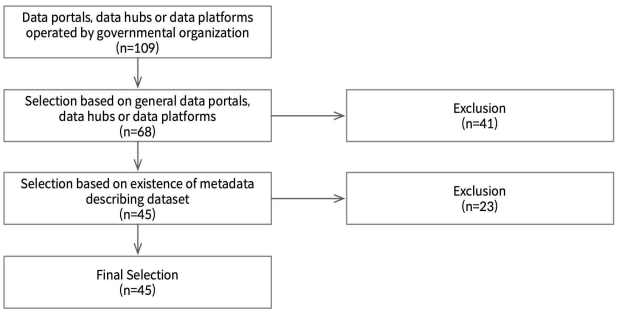

그림 1. 조사대상 선정과정
Fig. 1. Selection processes for survey

데이터 포털의 운영주체는 중앙부처 6개, 지방자치단체 13개, 공공기관 26개로 공공기관이 가장 많은 데이터 포털을 운영하고 있다. 대부분의 데이터 포털은 파일 데이터 (40개)와 API(28개)를 함께 제공하고 있다. 데이터 포털에 따라 제공하는 메타데이터 항목은 다르다. 데이터세트에 대한 메타데이터는 최소 5개에서 최대 47개이다. 그러나, 전체 데이터 포털에서 제공하는 메타데이터는 파일 데이터 15개, API 14개로 유형에 따른 차이는 크지 않다.

3-2 메타데이터의 추출과 선정

데이터세트의 메타데이터는 선정된 데이터 포털에서 제공하는 모든 항목을 대상으로 한다. 먼저 개별 데이터 포털의 메타데이터 항목을 추출하고, 정규화를 통해 비교 대상을 선정한다.

- 중복성 제거: 동일한 목적의 메타데이터 항목의 중복을 제거한다. 예를 들어, 경기데이터드림은 데이터세트에 대한 분류를 위해 '분류'와 '분류체계' 항목이 사용되고 있어 '분류체계'로 통일한다.
- 명시적인 항목명 부여: 데이터세트의 특정 항목에 대한 값이 있지만 항목명이 존재하지 않을 경우 (예: '데이터셋명', '제목'), 임의로 메타데이터 항목명을 부여한다.
- 유형에 따른 메타데이터: 데이터세트는 유형에 따라 다르게 정의한다. 예를 들어, 파일 데이터는 다운로드, 파일 크기와 같은 항목이 있고, API는 접근 URL, 인증정보, 활용신청건수와 같은 특징을 반영한다.

메타데이터 항목명과 값은 데이터 포털에 따라 다르다. 예를 들어, 데이터세트의 등록일자는 '등록일', '등록', '개방일', '최초 등록일자', 'DATA 개방일', '작성일', '공개일자', '최초 등록', '개방년월일', '데이터개방일', '데이터 최초 등록일', 'DATA 등록일', '생성날짜' 등 다양한 용어로 사용되고 있다. 한편, 메타데이터의 값은 일관된 형식이 존재하지 않는다. 공공데이터 공통표준용어는 용어명, 영문약어명, 저장형식, 표현형식 등 메타데이터의 값과 형식을 정의하고 있지만, 추출한 메타데이터에 적용되지 않고 있다.

표 1. 국내 데이터 포털의 메타데이터 개수
Table 1. The number of metadata elements provided by domestic data portals

Public Institution	Data Portal	URL	the number of metadata	
			FILE	API
Daegu Metropolitan City	D-Datahub	data.daegu.go.kr	13	13
Korea Water Resources Corporation	K-water Open Data Portal	opendata.kwater.or.kr	-	20
Korea Hydrographic and Oceanographic Agency	Korea Ocean Data Market Center	khoa.go.kr/komc	47	47
Ministry of Land, Infrastructure and Transport	Architecture Data Open System	open.eais.go.kr	6	-
Gyeonnggi Province	Gyeonggi Data Dream	data.gg.go.kr	15	17
Gyeonnggi Province	Gyeonggi Economic Portal	bigdata-region.kr	17	-
South Gyeongsang Province	Gyeongnam Bigdata Hub Platform	bigdata.gyeongnam.go.kr	18	14
Korea Expressway Corporation	Highway Public Data Portal	data.ex.co.kr	11	5
Ministry of the Interior and Safety	Public Data Portal	data.go.kr	27	19
Gwangyang City	Gwangyang City Public Data Portal	gwangyang.go.kr/data	12	-
Gwangju Metropolitan City	Gwangju Metropolitan City Bigdata Integration Platform	bigdata.gwangju.go.kr	12	8
Ministry of Land, Infrastructure and Transport	National Transport Information Center	its.go.kr	7	-
Korea Transport Institute	National Transport Data Open Market	bigdata-transportation.kr	20	-
National Library of Korea	National Bibliography LOD	lod.nl.go.kr	6	-
National Cancer Center	National Cancer Data Center	cancerdata.kr	-	10
Ministry of Land, Infrastructure and Transport	Data Integration Channel	data.molit.go.kr	17	-
Korea Meteorological Administration	Weather Data Open Portal	data.kma.go.kr	19	7
Korea Education and Research Information Service	Nice Education Information Open Portal	open.neis.go.kr	12	13
Ministry of Agriculture, Food and Rural Affairs	MAFRA Public Data Portal	data.mafra.go.kr	7	19
Korea Agro-Fisheries & Food Trade Corporation	KADX	kadx.co.kr	19	-
Korea Testing Laboratory	Digital Industry Innovation Bigdata Platform	bigdata-dx.kr	31	-
Korea Culture Information Service Agency	Culture Public Data Plaza	culture.go.kr/data/main/main.do	7	9
Korea Culture Information Service Agency	Culture Bigdata Platfform	bigdata-culture.kr	11	-
Korea Hydrographic and Oceanographic Agency	Ocean Data in Grid Framework	khoa.go.kr/oceangrid	-	6
Health Insurance Review & Assessment Service	Healthcare Bigdata Hub	opendata.hira.or.kr	18	15
Busan Metropolitan City	Busan Public Data Portal	data.busan.go.kr	19	15
Korea Forestry Promotion Institute	Forest Bigdata Exchange Platform	bigdata-forest.kr	16	-
Seoul Special Metropolitan City	Seoul Open Data Plaza	data.seoul.go.kr	17	16
National Fire Agency	Korea Fire Safety Bigdata Platform	bigdata-119.kr	21	-
Korean National Police University	Smart Policing Bigdata Platform	bigdata-policing.kr	14	-
Ministry of Food and Drug Safety	Food and Drug Data Portal	data.mfds.go.kr	9	-
National Cancer Center	CONNECT	bigdata-cancer.kr	-	17
Korea Social Security Information Service	Childcare Information Open Portal	info.childcare.go.kr	8	13
Incheon Metropolitan City	Incheon Public Data Portal	incheon.go.kr/data	18	7
South Jeolla Province	Jeollanam-do Bigdata Hub	data.jeonnam.go.kr	17	19
North Jeolla Province	Jeollabuk-do Bigdata Hub	bigdatahub.go.kr	20	16
Jeju Techno Park	Jeju Data Hub	jejudatahub.net	8	13
Public Procurement Service	Procurement Information Open Portal	data.g2b.go.kr	15	13
Changwon City	Changwon Bigdata Portal	bigdata.changwon.go.kr	15	-
North Chungcheong Province	Chungcheongbuk-do Bigdata Hub Platform	data.chungbuk.go.kr	18	-
Korean Intellectual Property Office	KIPRIS Plus	plus.kipris.or.kr	14	11
Anti-Corruption & Civil Rights Commission	Civil Complaint Bigdata	bigdata.epeople.go.kr	-	8
Ministry of Oceans and Fisheries	Ocean and Fisheries Bigdata Platform	vadahub.go.kr	21	-
Korea Maritime Institute	BigdataSea	bigdata-sea.kr	12	13
Korea Water Resources Corporation	Environment Bigdata Platform	bigdata-environment.kr	16	10

데이터세트의 등록일은 '2022-03-17', '2021-01-06 10:54:01', '2022.1.24.', '2021' 등 공통표준용어의 기준이 적용되지 않고 있다. 더불어, 데이터세트의 파일 크기는 서로 다른 크기의 단위(예: KB, MB)를 사용하고 있다.

Ⅳ. 공통 메타데이터의 선정

4-1 공통 메타데이터 선정

서로 다른 데이터 포털에서 데이터의 연계와 융합은 상호운용의 수준에 따라 구현 방법이 다르다. 메타데이터의 상호운용성은 데이터 포털과 데이터세트 수준에서 일치 여부를 비교할 수 있는 기준점이며, 메타데이터 항목명과 값을 공통적으로 규격화하고 합의를 통해 구현할 수 있다.

공통 메타데이터 모델은 다음의 기준을 따른다. 첫째, 2개 이상의 데이터 포털에서 사용하는 항목을 포함한다. 특정 데이터 포털에서 유일하게 제공되는 메타데이터 항목은 공통 메타데이터 모델에 포함하지 않는다(예: '공공포털 수집구분', '다운로드 미제공 사유', '자료생산 진행상태'). 둘째, 의미가 동일한 항목명은 표기명에 관계없이 동일한 메타데이터로 정의한다. 셋째, 항목명의 의미가 모호하거나 특정 데이터세트에 한정된 항목명은 공통 메타데이터 모델에 포함하지 않는다(예: '상품 구분').

추출된 메타데이터는 항목명을 기준으로 클러스터링을 수행하고, 이후 수작업으로 의미적 동일성을 판단한다. 그림 3은 공공데이터포털과 서울 열린데이터광장의 메타데이터를 매핑하고, 의미적 유사성을 기준으로 통일한 결과다. 예를 들어, '등록'과 '공개일자'는 데이터세트의 포털 등록일, '이용허락범위'와 '라이센스'는 데이터세트의 라이선스를 의미하는 동일한 항목으로 통일시킨다. 추출한 메타데이터는 총 135개이고, 데이터 포털을 위한 공통 메타데이터 모델은 총 45개 항목을 포함한다 (예: '데이터명', '등록일', '설명', '수정일', '제공기관', '확장자', '분류체계', '업데이트 주기', '조회수' 등).

4-2 공통 메타데이터의 수용력

공통 메타데이터는 개별 데이터 포털에서 공통적으로 적용되기 위해 적합한 수준의 대표성을 제공해야 한다. 공통 메타데이터의 수용력(Capacity; CCM)은 개별 데이터 포털이 제공하는 모든 메타데이터 항목(Total number of metadata items; TMpi)에 대한 공통 메타데이터의 일치(Identical number of common metadata items; ICMpi) 비율이다. 수용력은 0과 1사이의 값을 갖는다. 즉, 수용력이 1에 가까울수록 공통 메타데이터는 데이터 포털의 메타데이터와 일치 비율이 높다는 의미로 해석한다.

$$C_{CM} = \sum_{i=1}^{n} \frac{ICM_{p_i}}{TM_{p_i}}, \ (0 \leq C_{CM} \leq 1) \quad (1)$$

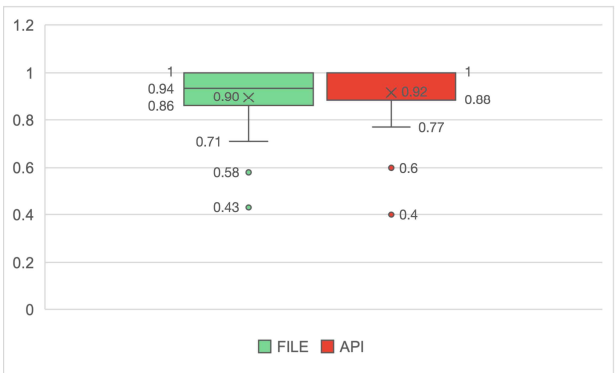

그림 2. 수용력(Capacity) 지수
Fig. 2. Capacity index of the common metadata

그림 2에서 보듯이, 선정된 모든 데이터 포털의 파일 데이터와 API의 수용력 지수는 각각 0.9, 0.92이다. 사분위수 범위(Inter Quartile Range, IQR)는 파일 데이터와 API가 각각 0.14와 0.11로 대부분의 데이터 포털이 유효한 범위에 포함된다. 이상치로 나타난 데이터 포털은 기상자료개방포털(파일0.58), 국가암데이터센터(API 0.6)와 KOMC 국가해양정보마켓센터(파일 0.43, API 0.4)다. 기상자료개방포털은 기상자료의 고유한 항목(예: '지점', '발표시각', '예측시각')을 제공하여 수용력 지수가 비교적 낮다. 국가암데이터센터와 KOMC 국가해양정보마켓센터는 자체적으로 포털에서 정의한 항목을 다수 포함하고 있어(예: '오퍼레이션 설명', '자료생산진행상태', '자동연계제공 여부', '자료처리수행 여부') 예외로 해석할 수 있다. 결과적으로 공통 메타데이터는 일반적인 데이터세트를 기술하는 메타데이터 집합을 포함하고 있다고 해석할 수 있다.

Ⅴ. DCAT-AP-KR 어휘

애플리케이션 프로파일은 특정한 응용 프로그램의 요구 사항을 충족하기 위해 추가된 제약 조건, 정책, 가이드라인과 함께 여러 메타데이터에서 선택한 용어를 사용하는 메타데이터 설계 명세(specification)다. 일반적으로 데이터 포털에서 제공하는 데이터세트는 DCAT 어휘로 기술할 수 있다. DCAT-AP는 데이터 포털에서 사용하는 DCAT의 애플리케이션 프로파일이다. 국내 데이터 포털에서 공통적으로 사용하는 어휘는 DCAT-AP와 구분되어 별도의 이름 체계를 부여하는 것이 바람직하다. 이런 의미에서 DCAT-AP-KR은 국내의 데이터 포털에서 공통적으로 사용할 수 있는 애플리케이션 프로파일이고, 이상적으로 RDF에 정의된 어휘를 기반으로 기술하거나, 이와 호환된다.

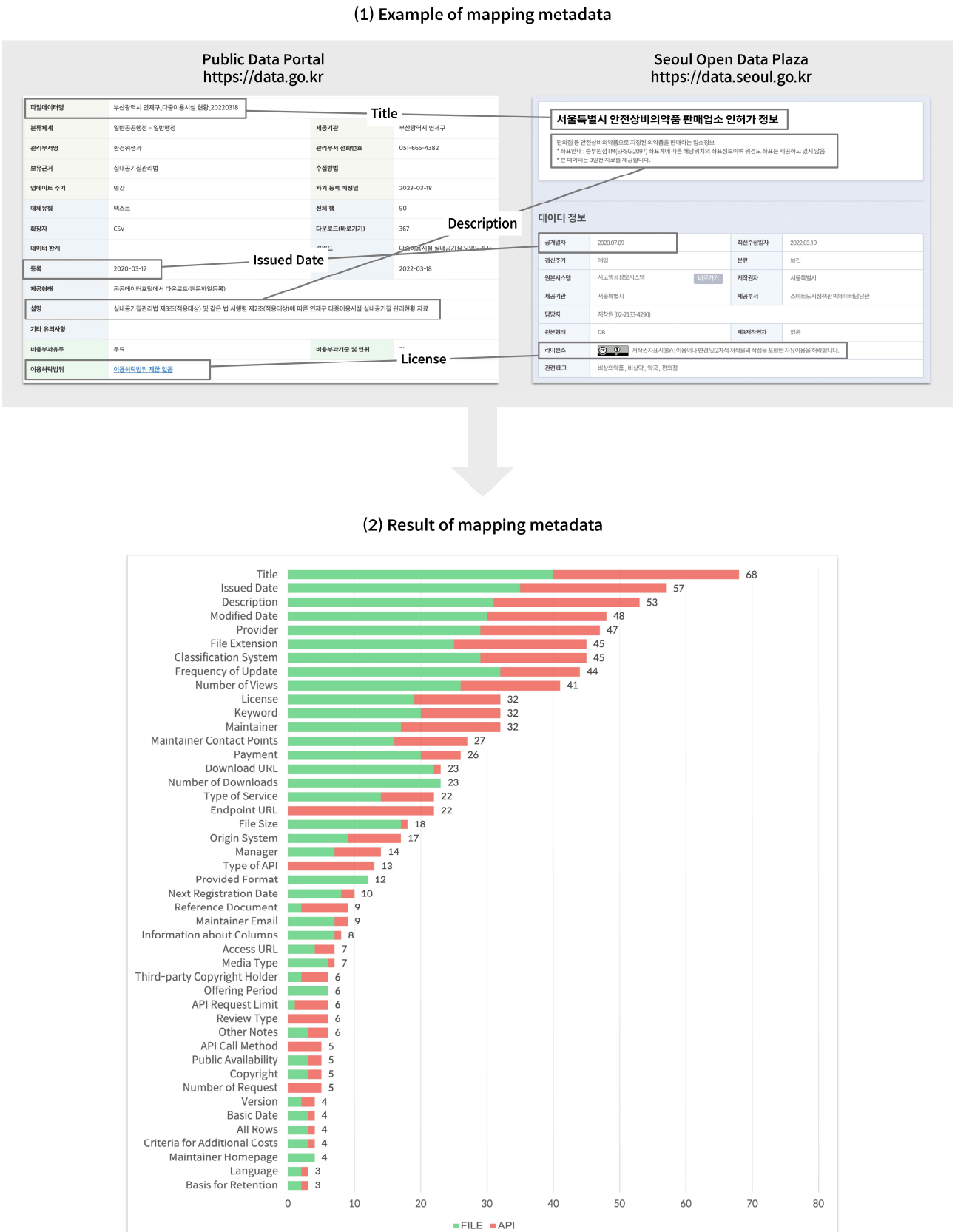

그림 3. 이종의 데이터 포털 메타데이터 매핑 예시와 메타데이터 매핑 결과
Fig. 3. The result of mapping metadata in two data portals

DCAT-AP-KR은 공통 메타데이터 어휘를 표현하는 데 목표가 있다. 데이터 포털에서 사용하는 메타데이터는 DCAT-AP 어휘를 적용하여 표현할 수 있지만, 국내의 데이터 포털에서만 사용되거나, 데이터 값의 형식 또는 의미가 다른 메타데이터는 별도의 어휘를 구성해야 한다. DCAT-AP-KR은 데이터세트 수준의 상호운용을 지원하고, 자유롭게 어휘를 확장할 수 있다.

5-1 모델링 원칙과 네임스페이스

DCAT-AP-KR은 DCAT-AP(DCAT Application Profile) 버전 2.1.0[7]의 의미론(semantics)를 준용하고, 최소의 신규 어휘를 정의한다. DCAT-AP-KR의 네임스페이스는 'http://vocab.datahub.kr/def/dcat-ap-kr/'이다. DCAT-AP-KR의 모델링 원칙은 다음과 같다.

- DCAT-AP-KR은 DCAT과 DCMI(Dublin Core Metadata Initiative)를 포함한 기존 RDF 어휘의 재사용을 원칙으로 한다. 예를 들어, 데이터세트와 데이터 서비스는 dcat:Dataset과 dcat:DataService 클래스와 관련 속성을 재사용한다.
- DCAT-AP-KR은 DCAT-AP 버전 2.1.0의 설계 원칙에 따라 필수 속성(mandatory property), 권장 속성(recommended property)과 선택 속성(optional property)을 정의한다. 필수 속성은 DCAT-AP-KR을 적용하면 반드시 기술해야 하는 메타데이터 항목이다. 권장 속성은 해당 속성에 대한 값이 존재할 경우 메타데이터의 적용을 권장하는 속성이다. 선택 속성은 해당 메타데이터 항목을 반드시 기술하지 않고 선택할 수 있다.
- DCAT-AP-KR은 공통 메타데이터 항목을 기준으로 정의한다. DCAT-AP 버전 2.1.0의 클래스와 속성을 따르고, 공통 메타데이터 항목을 표현하지 못할 경우, 신규 어휘를 정의한다. 메타데이터 항목의 값이 분류체계 또는 통제 어휘를 갖는 항목은 데이터 포털의 상황에 맞게 새롭게 정의할 수 있다.

4-2 DCAT-AP-KR 어휘의 지식 모델

DCAT-AP-KR의 스키마는 DCAT-AP의 구조와 의미를 그대로 따른다. 핵심 클래스는 데이터세트(dcat:Dataset), 배포(dcat:Distribution)와 데이터 서비스(dcat:DataService)로 구성된다. dcat:Dataset은 데이터의 집합(collection)으로, 하나 이상의 형식으로 접근가능하거나 다운로드 받을 수 있는 데이터세트를 의미한다. 데이터명, 설명, 제공기관, 분류체계, 키워드 등 데이터세트의 일반적인 특징은 dcat:Dataset 클래스의 속성으로 표현된다. dcat:Distribution은 데이터세트의 특정한 표현(a specific representation of a dataset)[3]을 의미하는 클래스다. 데이터세트는 다양한 방식으로 직렬화(serialize)할 수 있다.

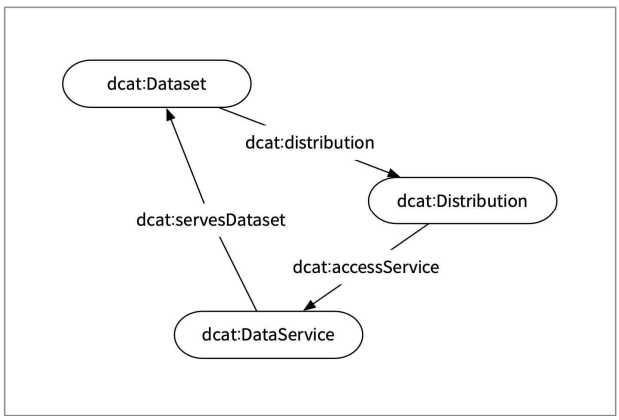

그림 4. DCAT-AP-KR의 핵심 클래스
Fig. 4. Core classes of DCAT-AP-KR

예를 들어, 동일한 데이터세트는 CSV, XLSX, JSON 등 서로 다른 데이터 포맷으로 제공할 수 있다. DCAT에서 이 관계는 dcat:distribution 속성으로 표현한다. 그림 4에서 보듯이, 어떤 데이터세트는 서로 다른 배포 형식이 존재할 수 있으면, 개별 형식은 dcat:Distribution의 인스턴스로 표현한다.데이터세트의 배포와 관련된 기술(예: 접근 URL, 이용허락범위, 다운로드 URL)은 dcat:Distribution의 속성으로 표현된다. dcat:DataService는 하나 이상의 데이터세트나 데이터 처리 기능에 접근할 수 있는 오퍼레이션(operation) 집합[3]이다. 예를 들어, API의 엔드포인트 URL과 기술문서는 dcat:DataService의 속성으로 표현될 수 있다. 데이터 서비스가 배포하는 데이터세트는 dcat:servesDataset으로 연결된다.

4-3 공통 메타데이터와 DCAT-AP-KR의 매핑

공통 메타데이터 항목은 DCAT-AP-KR의 속성(property)으로 표현될 수 있다. 공통 메타데이터 모델의 항목은 의미적으로 동일하거나 유사한 DCAT-AP의 속성을 재사용한다. 속성의 의미가 포괄적인 경우, 여러 개의 항목이 속성과 매핑될 수 있다. 예를 들어, '설명', '데이터 컬럼 정보', '기타 유의사항', '제공형태' 항목은 dct:description 속성과 매핑한다. DCAT-AP로 표현될 수 없는 항목은 외부 어휘를 재사용하거나 새로운 어휘로 정의한다. DCAT-AP-KR에서 추가적으로 정의된 속성은 항목의 중요도에 따라 우선순위를 부여한다. 공통 메타데이터의 45개 항목 중 31개 항목은 DCAT-AP의 속성을 재사용하고, 14개의 항목은 추가적으로 정의된다.

한편, DCAT-AP-KR의 각 속성의 값(rdfs:domain)으로 구체적인 값을 지정할 수 있다. DCAT-AP를 재사용하는 속성은 공역을 따르되, 국내 데이터 포털에 맞게 일부 통제어휘를 수정한다. 수정된 통제어휘는 총 5개로, 분류체계와 서비스유형, 이용허락범위, 확장자(매체유형), API 유형에 해당되는 값이다. 예를 들어, 이용허락범위는 공공누리 제1-4유형을 추가하고, 매체유형과 확장자는 한글(.hwp)을 추가한다.

표 2. 공통 메타데이터 항목과 DCAT-AP-KR의 매핑 결과
Table 2. Mapping results of both common metadata items and DCAT-AP-KR

Common metadata items	Priority	Domain	Property	Range	Additionally defined properties	Modified controlled vocabularies
Title	mandatory	dcat:Dataset	dct:title	rdfs:Literal		
	optional	dcat:Distribution				
	mandatory	dcat:DataService				
Description; Column Information; Other Notes; Provided Format	mandatory	dcat:Dataset	dct:description	rdfs:Literal		
	recommended	dcat:Distribution				
	optional	dcat:DataService				
Provider	recommended	dcat:Dataset	dct:publisher	koor:Organization		
Classification System	recommended	dcat:Dataset	dcat:theme	DCAT-AP-KR Controlled Vocabulary Dataset Theme		O
Maintainer; Manager	recommended	dcat:Dataset	dcatkr:maintainer	koor:Organization	O	
Keyword	recommended	dcat:Dataset	dcat:keyword	rdfs:Literal		
Maintainer Contact Points; Maintainer Email	recommended	dcat:Dataset	dct:contactPoint	vcard:Kind		
Basic Date	recommended	dcat:Dataset	dct:temporal	dct:PeriodOfTime		
Frequency of Update	optional	dcat:Dataset	dct:accrualPeriodicity	EU Vocabularies Frequency		
Number of Views	optional	dcat:Dataset	dcatkr:numberOfView	xsd:nonNegativeInteger	O	
Issued Date	optional	dcat:Dataset	dct:issued	xsd:date; xsd:dateTime		
	optional	dcat:Distribution				
Modified Date	optional	dcat:Dataset	dct:modified	xsd:date; xsd:dateTime		
	optional	dct:Distribution				
Type of Service	optional	dcat:Dataset	dct:type	DCAT-AP-KR Controlled Vocabulary Service Type		O
Origin System	optional	dcat:Dataset	dcatkr:derivedSystem	rdfs:Resource	O	
Public Availability	optional	dcat:Dataset	dct:accessRights	dct:RightsStatement		
Basis for Retention	optional	dcat:Dataset	dcatkr:legalBasis	rdfs:Resource	O	
Language	optional	dcat:Dataset	dct:language	EU Vocabularies Languages		
Version	optional	dcat:Dataset	owl:versionInfo	rdfs:Literal		
Payment	optional	dcat:Dataset	dcatkr:fee	xsd:boolean	O	
Criteria for Additional Costs	optional	dcat:Dataset	schema:offers	schema:Offer	O	
Next Registration Date	optional	dcat:Distribution	dcatkr:nextRegistrationDate	xsd:date; xsd:dateTime	O	
Access URL	mandatory	dcat:Distribution	dcat:accessURL	rdfs:Resource		
License	recommended	dcat:Distribution	dct:license	DCAT-AP-KR Controlled Vocabulary License		O
	optional	dcat:DataService				
File Extension; Media Type	recommended	dcat:Distribution	dct:format	IANA Media Types		O
Offering Period	recommended	dcat:Distribution	dcatap:availability	EU Vocabularies Planned Availability		
Number of Downloads	optional	dcat:Distribution	dcatkr:numberOfDownload	xsd:nonNegativeInteger	O	
Download URL	optional	dcat:Distribution	dcat:downloadURL	rdfs:Resource		
File Size	optional	dcat:Distribution	dcat:byteSize	xsd:decimal		
All Rows	optional	dcat:Distribution	dcatkr:numberOfRow	xsd:nonNegativeInteger	O	
Copyright; Third-party Copyright Holder	optional	dcat:Distribution	dct:rights	dct:RightsStatement		
Endpoint URL	mandatory	dcat:DataService	dcat:endpointURL	rdfs:Resource		
Reference Document; Review Type; API Call Method	recommended	dcat:DataService	dcat:endpointDescription	rdfs:Resource		
API Type	optional	dcat:DataService	dct:type	DCAT-AP-KR Controlled Vocabulary API Type	O	O
Number of Request	optional	dcat:DataService	dcatkr:numberOfRequest	xsd:nonNegativeInteger	O	
Request Limit	optional	dcat:DataService	dcatkr:numberOfRequestLimit	xsd:nonNegativeInteger	O	
Maintainer Homepage	optional	foaf:Agent	foaf:page	foaf:Document	O	

표 2는 공통 메타데이터 모델과 DCAT-AP를 매핑한 결과다. DCAT-AP-KR은 국내 데이터 포털의 데이터세트 기술을 위한 일관적인 속성과 값을 제공할 수 있고, 데이터세트의 기술이 의미적 수준에서 상호운용성이 확보된다.

V. 결 론

본 논문은 국내 데이터 포털에 공통으로 적용할 수 있는 메타데이터 요소 집합인 DCAT-AP-KR을 제안했다. DCAT-AP-KR은 DCAT의 애플리케이션 프로파일인 DCAT-AP에 국내의 데이터 포털에서 사용하는 어휘를 추가한 어휘 명세다. 공통 메타데이터는 공공기관이 운영하는 45개 데이터 포털의 135개 메타데이터를 추출했고, 의미 분석을 통해 45개 메타데이터를 공통 항목으로 선정했다. 수집대상의 데이터 포털에 대한 수용성 분석에 따르면, 공통 메타데이터의 수용성 지수는 약 90% 이상으로 공통 요소가 적절히 포함되었다. DCAT-AP-KR은 17개 속성이 추가·변형되었고, 개별 어휘의 값을 구체적으로 정의하고 있다. 예를 들어, 제공기관 속성은 행정기관을 값으로 정의하고 있고, 이를 위한 행정기관 어휘를 권고하고 있다. 즉, DCAT-AP-KR은 신규 어휘를 정의하며 동시에 기존의 어휘와 상호운용하는 방안을 함께 제공하고 있다.

연구 결과에 따르면, 국내의 데이터 포털에서 사용하는 메타데이터는 매우 다양하지만, 실제 사용되는 항목은 45개 정도다. 그러나, '등록', '등록일', '등록일자' 등 동일한 의미를 갖는 항목에 대한 다양한 어휘명이 존재하기 때문에 표준을 통해 어휘명을 통일하는 것이 필요하다. DCAT-AP-KR은 데이터 포털에서 사용하는 메타데이터의 항목과 값을 일관되게 표현하는 방법을 제공하기 때문에, 서로 다른 포털 사이의 의미적 연계를 가능하게 하고, 동시에 분산된 데이터 포털에 있는 데이터세트의 탐색을 지원할 수 있다.

그러나, 표준 어휘의 개발과 적용이 데이터 포털 또는 데이터세트의 연계와 융합을 보장하지 않는다. 표준(standards)은 표준화를 위한 합리적 기준으로, 합의에 의해 작성되고 인정된 기관에 의해 승인된 문서다. 여기서 합의(consensus)는 실질적인 문제에 대한 지속적인 반대가 없고 모든 관련 이해관계자의 견해를 고려하고 상충되는 주장을 조정하려는 과정을 포함한다. 즉, 표준은 일반적인 합의가 담긴 문서이고, 표준화는 해당 표준을 지속가능한 상태로 관련 생태계에 적용하고 발전시키는 일련의 활동이다. DCAP-AP-KR은 현재 정보통신단체표준으로 심의중이지만, 표준 제정 이후 어휘를 적용하고 활용하기 위한 소프트웨어의 개발과 지원을 제공해야 한다. 이런 맥락에서 DCAT-AP-KR를 처리하기 위한 라이브러리의 개발, 새로운 어휘를 추가하기 위한 규칙, DCAT-AP-KR과 기존의 RDF 어휘의 상호운용 방안에 대한 검증이 향후 연구되어야 한다.

참고문헌

[1] Open Data Strategy Council. Implementation plan for public data provision and utilization activation in 2022 [Internet]. Available: https://bit.ly/3U7eq0P

[2] Ministry of the Interior and Safety. Common standard terms for public data (Notification No. 2022-66, 2020.12.10.) [Internet]. Available: https://bit.ly/3VlOxf4

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

[4] R. V. Guha, D. Brickley and S. MacBeth, "Schema.org: Evolution of Structured Data on the web," *ACM Queue*, Vol. 13, No. 9, pp. 1-28, December, 2015.

[5] R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, P. Winstanley. Data Catalog Vocabulary (DCAT) - Version 3 [Internet]. Available: https://www.w3.org/TR/vocab-dcat-3/

[6] European Commission. About DCAT Application Profile for data portals in Europe [Internet]. Available: https://bit.ly/3g3jfJI

[7] ISA Programme of the European Commission. DCAT Application Profile for data portals in Europe Version 2.1.0 [Internet]. Available: https://bit.ly/3SZ9o6m

[8] M. Dekkers, S. Kotoglou, C. Nelson, M. Pellegrino, N. Hohn and V. Peristeras, "StatDCAT-AP, A Common Layer for the Exchange of Statistical Metadata in Open Data Portals," in *Proceeding of the 4th SemStats@ISWC*, Kobe, pp. 1-12, 2016. http://ceur-ws.org/Vol-1654/article-05.pdf

[9] P. Andrea, C. Vlado, F. Anders and L. Michael, "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, pp. 1-7, 2017.

[10] M. Cochez, N. Karim and I. Dimitriadis, Analysis of the DCAT-AP extensions, ISA Programme of the European Commission, SC353DI07171, pp. 1-31, 2017.

[11] DCAT-AP.de. Vocabulary and documents for DCAT-AP.de [Internet]. Available: https://www.dcat-ap.de/def/

[12] Open Knowledge Belgium. DCAT-BE Linking data portals across Belgium [Internet]. Available: http://dcat.be/

[13] Norwegian Digitalisation Agency. Standard for description of datasets, data services and data catalogs (DCAT-AP-NO) [Internet]. Available: https://bit.ly/3SQX9If

[14] ECH. DCAT-AP CH – The Swiss Application Profile for Data Portals and Catalogues [Internet]. Available: https://www.dcat-ap.ch/#
[15] Agency for Digital Italy. DCAT-AP_IT v1.1 – Italian profile of DCAT-AP [Internet]. Available: https://www.dati.gov.it/content/dcat-ap-it-v10-profilo-italiano-dcat-ap-0
[16] Publications Office of the European Union. The official portal for European data [Internet]. Available: https://data.europa.eu/en
[17] J. Klimek, "DCAT-AP representation of Czech National Open Data Catalog and its impact," *Journal of Web Semantics,* Vol. 55, pp. 69-85, March 2019. https://doi.org/10.1016/j.websem.2018.11.001
[18] F. Kirstein, B. Dittwald, S. Dutkowski, Y. Glikman and M. Hauswirth, "Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP," *International Conference on Electronic Government*, San Benedetto del Tronto, July 2019. https://doi.org/10.1007/978-3-030-27325-5_15
[19] W. Carrara, M. Dekkers, B. Dittwald, S. Dutkowski, Y. Glikman, F. Kirstein, N. Loutas, V. Peristeras and B. Wyns, Towards an open government data ecosystem in Europe using common standards, International Hellenic University, Thessaloniki, Greece, June 2017.
[20] Ministry of Science and ICT. Expansion and reorganization of 'Bigdata Map' … Easy check for 290,000 data [Internet]. Available: https://bit.ly/3Fs0dYf
[21] Ministry of Science and ICT. The nationwide knowledge platform 'Digital Jiphyenonjeon' is fully promoted [Internet]. Available: https://bit.ly/3FusJZ1
[22] C. Song, and H. Kim, "Improvements of public data policy through data portal analysis of local governments," *Journal of Digital Contents Society*, Vol. 23, No. 4, pp. 697-705, April 2022. https://doi.org/10.9728/dcs.2022.23.4.697
[23] K. Park, T. Kim, N. Amin, H. Seo, H. Kim, H. Won and Y. Lee, "Design and Implementation of Framework based on DCAT-AP for CKAN," *Database Research*, Vol. 33, No. 3, pp. 40-51, 2017.
[24] D. K. Shin, S. H. Lee, J. Kang and E. M. Park, "Data Catalogue Standards Based on DCAT for Transportation Data: DCAT-Trans," Journal of Korean Society of Transportation, Vol. 37, No. 5, pp. 430-444, October 2019. https://doi.org/10.7470/jkst.2019.37.5.430
[25] J. Kim, H. Yoon, Y. Kwon and S. T. Kim, "Metadata Design for Ecological Research Data: Focused on DCAT," *Korean Library And Information Science Society*, Vol. 51, No. 4, pp. 249-278, 2020. https://doi.org/10.16981/kliss.51.4.202012.249
[26] J. Park, K. Kim, S. Kim, and J. Youn, "Transformation Method for Publishing DCAT based Metadata in Data Repository on Web," in *Proceeding of the Korea Information Processing Society Conference*, Seoul: SEL, pp. 291-493, 2021. https://doi.org/10.3745/PKIPS.y2021m11a.491
[27] E. M. Park, and J. H. Kang, "Developing RDF Meta data Graph for Transportation Open Data Platform," *Journal of Korea institute of intelligent transport systems*, Vol. 20, No. 6, pp. 110-116, December 2021. https://doi.org/10.12815/kits.2021.20.6.110
[28] E. W. Woo, H. Won, and M. C. Nguyrn, "W3C DCAT·PROF Recommendation and Datamap Structure Design and Implementation," in *Proceedings of the Korean Information Science Society Conference*, Seoul: SEL, pp. 774-776, 2022.
[29] K. Park, H. Won and K. H. Ryu, "A Design and Implementation of a DCAT-based Metadata Transformation Tool for Interoperability in Open Data Platforms," *Journal of Digital Contents Society*, Vol. 19, No. 1, pp. 59-65, January 2018. https://doi.org/10.9728/dcs.2018.19.1.59

박하람(Haram Park)

2021년 : 중앙대학교 사회학과 (문학사)

2017년~2021년: 중앙대학교 사회학과
2021년~현 재: 중앙대학교 문헌정보학과 정보학 석사과정
※관심분야 : 지식그래프, 메타데이터, 공공데이터 등

김학래(Haklae Kim)

2010년 : 아일랜드 국립대학교 (공학박사)

2004년~2009년: Digital Enterprise Research Institute, Ireland
2009년~2016년: 삼성전자
2017년~2019년: 한국과학기술정보연구원
2019년~현 재: 중앙대학교 문헌정보학과 교수
※관심분야 : 지식그래프, 인공지능, 데이터 사이언스 등