문화예술기관 기본정보의 품질개선과 연계를 위한 지식그래프 구축*

Constructing a Knowledge Graph for Improving Quality and Interlinking Basic Information of Cultural and Artistic Institutions

선은택 (Euntaek Seon)**
김학래 (Haklae Kim)***

초 록

정보통신 기술이 빠르게 발전하면서 데이터의 생산 속도가 급증하였고, 이는 빅데이터라는 개념으로 대표되고 있다. 단시간에 데이터 규모가 급격하게 증가한 빅데이터에 대해 품질과 신뢰성에 대한 논의도 진행되고 있다. 반면 스몰데이터는 품질이 우수한 최소한의 데이터로, 특정 문제 상황에 필요한 데이터를 의미한다. 문화예술 분야는 다양한 유형과 주제의 데이터가 존재하며 빅데이터 기술을 활용한 연구가 진행되고 있다. 하지만 문화예술기관의 기본정보가 정확하게 제공되고 활용되는지를 탐색한 연구는 부족하다. 기관의 기본정보는 대부분의 빅데이터 분석에서 사용하는 필수적인 근거일 수 있고, 기관을 식별하기 위한 출발점이 된다. 본 연구는 문화예술 기관의 기본정보를 다루는 데이터를 수집하여 공통 메타데이터를 정의하고, 공통 메타데이터를 중심으로 기관을 연계하는 지식그래프 형태로 스몰데이터를 구축하였다. 이는 통합적으로 문화예술기관의 유형과 특징을 탐색할 수 있는 방안이 될 수 있다.

ABSTRACT

With the rapid development of information and communication technology, the speed of data production has increased rapidly, and this is represented by the concept of big data. Discussions on quality and reliability are also underway for big data whose data scale has rapidly increased in a short period of time. On the other hand, small data is minimal data of excellent quality and means data necessary for a specific problem situation. In the field of culture and arts, data of various types and topics exist, and research using big data technology is being conducted. However, research on whether basic information about culture and arts institutions is accurately provided and utilized is insufficient. The basic information of an institution can be an essential basis used in most big data analysis and becomes a starting point for identifying an institution. This study collected data dealing with the basic information of culture and arts institutions to define common metadata and constructed small data in the form of a knowledge graph linking institutions around common metadata. This can be a way to explore the types and characteristics of culture and arts institutions in an integrated way.

키워드: 문화예술, 스몰데이터, 지식그래프, 시맨틱웹, 링크드 데이터
arts & culture, small data, knowledge graph, semantic web, linked data

---

* 이 논문은 2022년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.
** 중앙대학교 일반대학원 문헌정보학과 정보학전공 석사과정(euntaekseon@gmail.com) (제1저자)
*** 중앙대학교 사회과학대학 문헌정보학과 교수(haklaekim@cau.ac.kr) (교신저자)
■ 논문접수일자 : 2023년 11월 20일 ■ 최초심사일자 : 2023년 12월 8일 ■ 게재확정일자 : 2023년 12월 11일
■ 정보관리학회지, 40(4), 329-349, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.4.329

1. 서 론

1.1 연구의 필요성과 목적

데이터는 '21세기의 새로운 원유'라는 비유에서 보듯이 우리의 일상에서 보편화되고 있다. 휴대폰을 통해 전달되는 하나하나의 공지정보, 소셜 미디어에서 업로드되는 정보, 버스요금을 지불하는 데 사용하는 센서정보, 도서관에서 책을 대출하는 정보 등 일상의 모든 활동이 데이터로 생산되고 유통된다. 이러한 흐름은 빅데이터로 대표된다. 빅데이터는 일반적인 데이터베이스의 관리와 분석체계로 감당하기 어려운 방대한 규모의 데이터다(복경수, 유재수, 2017). 인류의 사회문화적 환경이 정보통신과 밀결합되고 그 속도가 가속화됨에 따라 빅데이터에 대한 개념은 규모(volume), 속도(velocity), 다양성(variety)을 넘어 가치(value), 정확성(veracity), 가변성(variability)과 같은 새로운 특징을 포함하고 있다.

특히, 데이터의 정확성은 빅데이터의 품질과 활용 측면에서 매우 중요한 특징이다. 단시간에 데이터 규모가 급격하게 증가한만큼 데이터의 품질과 신뢰성에 대한 논의가 다양하게 진행되고 있다(김학래, 2017). 데이터 품질의 평가는 데이터의 활용에 있어 중요한 요소이지만, 데이터의 유형과 값은 하나의 기준으로 정립되기 어렵고, 빅데이터를 대상으로 데이터의 정확성을 측정하는 것이 현실적으로 어려움이 있는 것이 사실이다. 반면 스몰데이터(small data)는 품질이 우수한 최소한의 데이터를 의미하고(Latzko-Toth, Bonneau, & Millette, 2017), 종종 빅데이터와 다른 특성으로 구분해서 설명된다. 스몰 데이터의는 규모가 작은 데이터가 아니라, 정해진 문제 상황에 필요한 데이터이고 고품질인 상태를 갖고 있다는 특징이 있다. 이런 측면에서, 스몰 데이터는 지속적인 재사용이 가능하고 다른 데이터와 연계하여 활용할 수 있는 기반으로 인식되고 있다. 예를 들어, 온라인 쇼핑몰은 고객, 상품, 판매업체 등 다양한 정보와 이들 사이의 활동을 실시간으로 관리한다. 한 명의 고객이 상품을 탐색하고 장바구니에 등록하고, 상품을 구입하는 일련의 활동이 반복될수록 데이터는 빅데이터의 특성을 갖는다. 반면, 고객이 갖고있는 기본정보(성명, 전화번호, 주소, 결제정보 등)는 매우 정확한 값을 갖고, 고객의 모든 데이터 활동에 반복적으로 활용되는 필수정보다. 즉, 스몰 데이터는 빅데이터와 비교하면 규모가 작을 수 있지만, 데이터의 정확성과 재사용이 높은 특징을 갖는다. 이런 측면에서 스몰 데이터의 중요성이 강조되고 있고, 데이터 분석과 활용을 위해 스몰 데이터에 대한 연구의 필요성이 제기되고 있다(Strickland, 2022).

문화예술 분야는 서지데이터, 소장데이터, 연구데이터, 장서데이터, 인용데이터 등 다양한 유형과 주제에 대한 데이터가 존재하고, 빅데이터 기술을 활용한 연구가 진행되고 있다. 그러나, 기관의 기본 정보가 정확하게 제공되고 활용되는지를 탐색한 연구는 부족한 것이 현실이다. 문화예술 분야의 기관 정보는 공공데이터 형식으로 개방되고 있지만, 메타데이터의 규격, 데이터 값의 일관성이 통일되어 있지 않다. 스몰 데이터 관점에서 보면, 기관의 기본정보는 대부분의 빅데이터 분석에서 사용하는 필수적인 근거일 수 있고, 기관을 식별하기 위한 출발점이 되는 중요

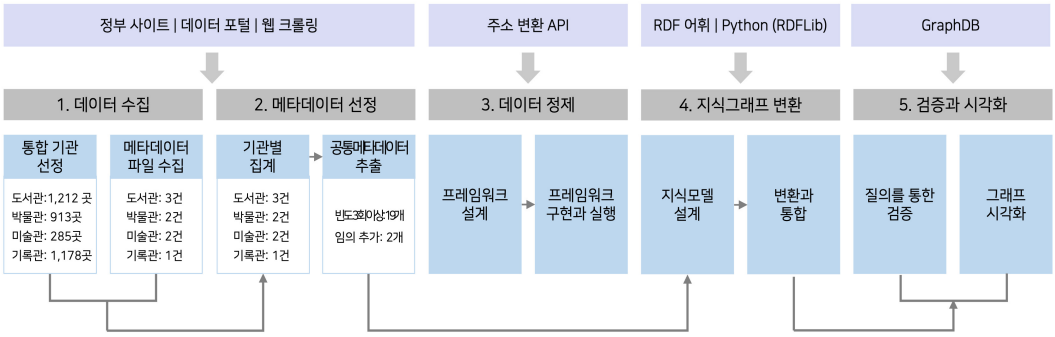
한 데이터다.

본 연구는 국내 문화예술 관련 기관의 기본정보에 대한 데이터를 조사하고, 공통적인 데이터 제공을 위한 메타데이터 요소를 정의한다. 규격화된 기본정보는 기관에 대한 데이터를 표현하는 기준이 되며, 동시에 기관의 다양한 정보를 연결하는 데 사용될 수 있다. 본 연구는 법률적 기준에 근거하여 문화예술기관에 대한 목록을 정의하고, 공공데이터로 공개된 기관정보를 수집하여 메타데이터 요소를 정의한다. 특히, 기관의 기본정보를 지식그래프로 구축하여 기관 사이의 연계 관계를 정의하고, 기관의 유형, 특징을 탐색할 수 있는 방안을 제공한다.

1.2 연구방법

본 연구는 문화예술기관의 기본정보를 수집하고, 지식그래프로 구축하는 것으로 목표로 한다. 따라서 본 연구는 문화예술기관 목록의 정의, 지식그래프 구축을 위한 메타데이터의 설계와 변환 과정을 포함하고 있다. 〈그림 1〉에서 보듯이, 연구 단계는 데이터 수집, 메타데이터 선정, 데이터 정제, 지식그래프 변환, 검증과 시각화를 포함하고 있다.

1) 1단계 (데이터 수집): 통합할 데이터를 수집하는 것으로 도서관, 박물관, 기록관, 미술관의 각 현황 데이터를 수집한다. 이 단계는 목적에 따라 두 과정으로 나뉜다. 첫 번째는 지식그래프로 통합할 기관을 선정하기 위한 데이터를 수집하는 과정이다. 이 데이터는 정부부처에서 공식적으로 개방한 데이터에 포함되어 있는 기관에 한정한다. 두 번째는 기관에 대한 메타데이터를 수집한다. 데이터포탈이나 플랫폼을 통해 개방된 각 기관의 현황 데이터를 수집하고, 필요하다면 웹에 있는 정보를 크롤링하여 데이터를 수집한다.
2) 2단계 (메타데이터 선정): 네 종류 기관의 데이터에서 공통적으로 사용하고 있는 공통 메타데이터 항목을 선정한다. 메타데이터 선정은 두 번 진행된다. 먼저 각 기관의 현황 데이터끼리 비교하는 것으로, 네 개의 기관별로 메타데이터 항목을 집계하여 기관별 한 벌의 메타데이터 세트를 만든다. 두 번째 작업은 기관별로 생



〈그림 1〉 연구과정 도식화

성된 한 벌의 메타데이터 세트인 총 네 개의 메타데이터 세트를 종합적으로 검토하는 과정으로, 4개의 기관 중 3개 이상의 기관에서 사용된 항목을 추출하고, 그 외 필요하다고 판단되는 항목도 추출한다.

3) 3단계 (데이터 정제): 2단계에서 선정한 공통 메타데이터 항목의 실제 값을 대상으로 정제를 진행한다. 정제 목적은 크게 두 가지다. 하나는 결측 데이터를 해결하는 것으로, 이용자가 데이터의 맥락을 훼손하지 않고 채울 수 있는 결측 데이터는 채운다. 또한 같은 정보를 나타내는 메타데이터라도 데이터별로 사용하는 인스턴스 형식이 일치하지 않을 수 있다. 따라서 메타데이터 유형별로 인스턴스 형식을 통일하는 정제도 같이 진행한다.
4) 4단계 (지식그래프 변환): 지식 모델 설계 후 그래프로 변환하는 과정이다. 선정한 공통 메타데이터 항목에 적합한 RDF 어휘와 속성을 부여하여 지식모델을 설계한다. 지식모델을 기반으로 실제로 지식그래프로 구축하여 네 종류 기관의 현황 데이터를 통합하는 과정이다.
5) 5단계 (검증): 구축한 지식그래프를 검증하고 활용하는 과정이다. 트리플 저장소인 그래프데이터베이스에 변환한 데이터를 저장한다. 그 후 SPARQL 쿼리를 이용하여 도서관, 박물관, 미술관, 기록관의 데이터를 통합적으로 탐색할 수 있는지 검증한다. 또한 지식그래프 시각화를 통해 상호 연결된 관계와 구조를 확인한다.

1.3 선행연구

1945년 7월, 바네바 부시(Vannevar Bush)는 최초로 데이터 사이의 연결을 소개한다. 바네바 부시는 Memex라는 가상의 기계를 통해 데이터를 저장할 수 있고, 이 데이터들을 서로 연결할 수 있다고 소개했다(Bush, 1945). 이 이론은 이후 테드 넬슨에 의해 하이퍼텍스트라는 이론으로 발전했고, 팀 버너스리(Tim Berners-Lee)는 하이퍼텍스트 이론을 웹에 적용하였다.

팀 버너스리는 1998년 시맨틱 웹이라는 개념을 제안했다(Berners-Lee, 1998). 웹 상에서 데이터에 의미적 관계를 부여할 수 있는 웹으로, 웹 페이지를 구성하는 데이터 간의 연결을 목표로 하여 데이터 중심의 웹을 강조한다(이용주, 2014). 즉 시맨틱 웹은 데이터 간의 의미 있는 연결 관계를 형성하여 사람과 기계가 데이터를 읽고 처리할 수 있는 환경을 의미한다(윤소영, 2013). 이때 링크드 데이터 기술은 기존의 데이터를 시맨틱 웹 환경에 맞게 변환하는 기법과 그 결과 데이터를 말한다. 구글은 링크드 데이터 기술을 검색 시스템에 도입하기 위한 방법으로 지식그래프를 발표하면서, 의미를 이용하여 검색하는 전략을 사용하기 시작했다.

지식그래프는 이기종 데이터 사이의 관계를 방향성 있는 그래프로 나타낸 것을 의미한다(Wang et al., 2017). 지식그래프는 개체의 상호 연결된 설명을 저장하는 그래프 형태의 저장된 지식을 의미한다(Fensel et al., 2020; Hogan et al., 2021). 지식그래프는 주체-객체 관계(subject-object relationship)를 사용하여 개체 간의 관계를 표현한다. 개체는 사람, 장소, 사물, 사건 등과 같은 실체를 나타내고, 관계는 개체

간의 의미적 관계를 나타낸다. 예를 들어, "중앙대학교"는 개체이고, "대한민국의 대학교"는 관계로 표현할 수 있다. 지식그래프는 구글이 상용 서비스를 시작하면서 사용되었는데, 이전에 학계를 중심으로 시맨틱 웹, 링크드 데이터라는 용어를 보편적으로 활용하였다(Tiwari, Al-Aswadi, & Gaurav, 2021). 시맨틱 웹은 컴퓨터가 이해할 수 있는 의미론적 데이터를 표현하거나 처리하는 웹 환경이고, 링크드 데이터는 RDF를 사용하여 표현된 데이터를 서로 연결하는 기술과 그 결과 데이터를 말한다. 지식그래프는 시맨틱 웹의 한 구성 요소이며, 링크드 데이터를 기반으로 구축된다. 지식그래프는 의미론적 데이터를 그래프 형태로 표현한 것으로, 링크드데이터를 사용하여 데이터 간의 관계를 연결한다(Abu-Salih, 2021; Gutiérrez & Sequeda, 2021). 지식그래프에서의 지식모델링은 특정한 주제에 대한 의미와 관계를 설계하는 과정을 포함한다. 지식모델은 도메인을 특성을 반영하여 새로운 어휘를 정의하거나 기존에 개발된 어휘를 목적에 맞게 재사용할 수 있다. 일반적으로 데이터의 의미적 상호운용성을 확보하기 위해, 지식모델과 어휘는 재사용을 권고한다(Grau et al., 2008).

대표적인 RDF 어휘 중 하나인 SKOS는 웹 온톨로지로서 분류체계의 개념, 관계, 속성을 정의하고, 이를 사용하여 분류체계를 표현하고 관리할 수 있도록 지원하는 어휘 표준이다(Miles & Pérez-Agüera, 2007). SKOS를 사용하여 표현된 분류체계는 웹에서 쉽게 공유되고 교환될 수 있다(Sanchez-Alonso & Garcia-Barriocanal, 2006). 도서관의 경우, SKOS를 사용하여 도서, 저자, 출판사, 장르, 주제 등과 같은 어휘를 표현할 수 있다. 즉, 서로 다른 도서관이 공통의 표준어휘를 적용하면, 다양한 데이터를 보다 효율적으로 관리하고, 사용자에게 보다 유용한 검색 결과를 제공할 수 있다.

Schema.org[1]는 2011년 주요 검색엔진인 Bing, Google, Yahoo 그리고 Yandex가 공동으로 개발한 어휘로, 웹에서 여러 주제(예: 기관, 장소, 사람, 물건 등)를 의미적으로 표현할 수 있다. 이 어휘는 웹 데이터 사이의 상호운용성을 위해 단일한 스키마를 제공하는 것을 목표로 생성되었다(Iliadis et al., 2022). 웹 페이지에 존재하는 정보를 구조화하여 검색엔진이 더욱 정확하게 분석할 수 있도록 도움을 주며, Alexa나 Google Assistant 같은 가상 비서에도 데이터를 공급한다(Kollar et al., 2018). 현재 Schema.org는 803개의 Type과 1466개의 Property 등을 제공하고 있다.

Basic Geo Vocabulary[2]는 위도와 경도의 사물에 대한 공간 데이터를 표현하기 위한 RDF 어휘로, WGS84 데이터를 사용한다. 이 어휘는 스키마를 변경할 필요없이 관련 RDF 어휘를 사용하여 지도뿐만 아니라 지도 위에 있는 엔티티도 설명이 가능하다는 장점이 있다. Organization Ontology[3]는 기관에 대해 기술할 수 있는 핵심 온톨로지 어휘로 도메인별 확장을 통해 조직과 기관의 분류를 추가할 수 있고, 조직의 활동과 같은 관련 정보들을 추가할 수 있다는 장점이 있다.

1) "Schema.org", https://schema.org//, 2023년 11월 20일 접속
2) "Basic Geo Vocabulary", https://www.w3.org/2003/01/geo/, 2023년 11월 20일 접속
3) "Organization Ontology", https://www.w3.org/TR/vocab-org/, 2023년 11월 20일 접속

데이터 조정(data reconciliation)은 서로 다른 데이터출처에서 동일한 개체를 나타내는 데이터를 통합하는 작업을 말한다(Mandal, 2022; Mongiovì et al., 2015). 데이터조정을 통해 서로 다른 데이터출처에서 중복되는 데이터를 제거하고, 정확하고 유용한 데이터를 생성할 수 있다. 일반적으로 데이터조정은 동일한 개체의 인식, 데이터의 통합, 중복 데이터 제거, 수동 또는 자동적인 조정을 포함한다(Monnin et al., 2019). 데이터 조정은 대규모 데이터에 있는 동일한 개체를 인식하여 데이터를 연계하거나 통합하는 핵심 기술이고, 최근 거대언어모델과 지식그래프 기술을 활용하는 방식으로 발전하고 있다(Daruna et al., 2022; Sohmen & Rossenova, 2022). 특히, 생성 인공지능의 문제로 지적되고 있는 환각(hallucination) 현상을 분석하고 오류를 수정하는 데 데이터조정 기술이 응용되고 있다.

본 연구는 국내 문화예술 관련 기관의 분산된 현황 데이터를 통합하기 위해 직접 기관 데이터를 수집하고 정제하여 지식그래프로 변환하는 방안을 제안한다. 수집한 데이터를 검토하여 공통 메타데이터 항목을 추출하고 지식그래프로 구현한다. 추출한 공통 메타데이터는 기존의 RDF 어휘를 재사용하고, 의미와 관계가 충분하지 않은 어휘는 신규로 정의한다.

2. 기관별 데이터의 수집과 분석

2.1 데이터 선정

도서관, 박물관, 미술관, 기록관의 현황 데이터를 수집하였다. 수집하는 데이터의 목적은 두 가지로 구분할 수 있다. 하나는 기관의 현황 수를 나타내기 위한 데이터로, 도서관, 박물관, 미술관, 기록관을 관할하는 상위 정부기관에서 공식적으로 개방한 단일 데이터를 수집하였다. 해당 데이터에서 제시한 기관을 중심으로 메타데이터를 연결하고 지식그래프로 구축하였다.

두 번째는 앞에서 수집한 데이터 외, 추가적으로 기관에 대한 정보를 나타낼 때 사용하는 메타데이터를 추출하기 위한 데이터다. 정부 기관 사이트 외에도 각 기관의 현황 데이터는 다양한 플랫폼을 통해 개방되기도 하는데, 이 데이터를 두 번째 단계에서 수집하였다. 그 후 첫 번째 단계에서 수집한 데이터와 두 번째 단계에서 수집한 데이터를 종합적으로 검토하여 기관의 현황을 나타낼 때 사용하는 메타데이터 항목을 추출하였다. 만약 정부기관에서 고시한 데이터 외 다른 포탈에서 데이터를 제공하지 않는다면 웹에 존재하는 정보를 크롤링하여 직접 두 번째 데이터 파일을 생성하였다.

2.2 데이터 수집 현황

국내 도서관, 박물관, 미술관의 현황 수는 문화체육관광부에서 2022년에 개방한 문화기반시설총람2022 데이터를 근거로 정의하였다. 문화기반시설총람은 전국의 문화기반시설에 대한 현황 정보를 하나의 파일에 정리하여 개방하는 데이터로 2003년부터 문화체육관광부에서 발간하고 있다. 도서관과 박물관, 미술관 외에도 생활문화센터, 문예회관 등에 대한 정보를 담고 있다. 문화기반시설총람2022를 개방한 날짜는 2022년 12월 29일이며, 데이터 기준 날

짜는 2022년 1월 1일이다. 총람에 따르면 국립 도서관은 4곳, 공공도서관은 1,208곳으로 도서관은 총 1,212곳이 있다. 박물관은 국립, 공립, 대학, 사립 박물관을 통틀어 전국에 913곳이 있고, 미술관은 국립, 공립, 대학, 사립 박물관을 통틀어 285곳이 있다.

도서관에 대한 메타데이터 통합용 데이터는 공공데이터포탈에서 개방하는 전국도서관표준데이터와 문화 빅데이터 플랫폼에서 개방하는 도서관 정보 데이터를 활용하였다. 표준데이터는 각 지방자치단체에서 제공한 데이터를 합하여 주기성 데이터로 개방하고 있다. 도서관 정보 데이터는 국립중앙도서관에서 제공한 데이터로 한 달에 한 번씩 업데이트된다.

박물관과 미술관에 대한 메타데이터 통합용 데이터는 공공데이터포탈을 통해 개방된 전국박물관미술관정보표준데이터로, 전국도서관표준데이터와 마찬가지로 지방자치단체에서 데이터를 제공하고 주기적으로 업데이트된다.

국내 기록관의 현황에 대해 포탈을 통해 테이블 형태의 데이터로 개방된 자료는 없다. 따라서 기록관에 해당하는 기관을 선정할 때 참고한 데이터는 행정안전부 국가기록원 고시 국가기록원관할 기록관을 대상으로 한정하였다. 행정안전부에서 2022년 7월 8일에 웹 사이트를 통해 개방하였으며, 전국의 1,304곳의 기록관이 국가기록원 관할에 해당된다.

기록관의 메타데이터는 웹에 존재하는 정보를 크롤링하였다. 사단법인 한국기록전문가협회는 2020년 12월 7일에 전국의 기록물 보유기관 현황을 구글 지도로 제작하여 베타버전을 오픈하였다.[4] 이 지도는 전국의 기록관을 소관기관에 따라 7개의 카테고리로 나누고, 각 기록관에 대한 기본적인 소개, 연락처, 좌표, 기록전문가 배치인원 등의 정보를 포함하고 있다.

2.3 메타데이터 분석

4-2장에서 기관별로 수집한 데이터에서 메타데이터를 추출하여 검토하였다. 여러 기관에서 공통적으로 사용되는 메타데이터 항목이 있는 반면, 특정 데이터에서만 사용하는 메타데

〈표 1〉 수집한 데이터 목록

구분	기관 선정용 데이터		메타데이터 통합용 데이터	
	데이터명	출처	데이터명	출처
도서관	전국 문화기반시설 총람 2022	문화체육관광부	전국도서관표준데이터	공공데이터포탈
			도서관 정보	문화빅데이터플랫폼
박물관	전국 문화기반시설 총람 2022	문화체육관광부	전국박물관미술관정보표준데이터	공공데이터포탈
미술관	전국 문화기반시설 총람 2022	문화체육관광부	전국박물관미술관정보표준데이터	공공데이터포탈
기록관	2022년 국가기록원 관할 기록관 설치 현황	행정안전부 국가기록원	전국기록관지도	사단법인 한국기록전문가협회

4) “전국기록관지도”, https://www.google.com/maps/d/u/0/viewer?mid=1mf_JrZkYijitsG83xfIdrErNUZCt3nxo&ll=-3.81666561775622e-14%2C132.5454629801115&z=1, 2023년 11월 20일 접속

이터 항목이 있었다. 또한 동일한 의미를 갖는 메타데이터지만 명칭이 다른 경우도 있었다. 메타데이터를 검토하여 공통 메타데이터를 추출하고, 공통 메타데이터를 중심으로 지식그래프를 구축하였다.

공통 메타데이터의 선정은 두 번 진행하였다. 1차 메타데이터 선정은 기관별로 메타데이터를 통합하는 과정으로, 각 기관별로 메타데이터의 중복이 없는 한 벌의 메타데이터 세트가 결과물로 생성된다. 각 데이터에서 사용하는 메타데이터명은 다르지만 의미가 동일한 경우가 있다. 동일한 의미의 메타데이터가 중복되어 포함되는 경우를 방지하기 위해 실제 메타데이터의 인스턴스 값을 참고하여 집계하였다. 만약 메타데이터명이 데이터별로 다르다면 메타데이터명을 일관적으로 통일하여 새로 부여하였다. 새로운 명칭은 미사여구를 붙이지 않고, 메타데이터의 의미가 통할 만큼 최소한의 명칭으로 간결하게 부여하였다. 〈표 2〉는 기관별 메타데이터의 중복을 제거하고 집계 한 후, 메타데이터 개수를 정리한 것이다.

1차 메타데이터 선정의 결과물을 바탕으로 2차 메타데이터 선정을 진행하였다. 네 벌의 메타데이터 세트 중, 메타데이터의 출현 빈도를 계산하여 3회 이상 사용된 항목을 추출하였다.

또한 검토하면서 필요하다고 판단되는 메타데이터 2개를 추가적으로 추출하였다. 첫 번째는 기록관의 유형을 나타내기 위해 필수적이라고 판단되는 '기관유형' 메타데이터다. 도서관과 박물관, 미술관의 경우 기관을 설립한 주체의 유형에 따라 해당 기관의 유형을 나누며, 실제 입력되는 인스턴스 값도 세 개의 기관이 매우 유사하다. 도서관법 4조에 따르면 도서관은 설립・운영 주체에 따라 국립, 공립, 사립으로 구분한다(도서관법. 법률 제19592호). 국립은 국가가 설립하는 기관, 공립은 지방자치단체 또는 교육감이 설립하는 도서관, 사립은 민법이나 상법 등에 따라 설립된 법인・단체・개인이 설립한 도서관을 의미한다. 박물관 및 미술관 진흥법 3조에 따르면 도서관법과 마찬가지로 박물관과 미술관은 국립, 공립, 사립으로 나누고 추가적으로 대학으로도 구분이 된다(박물관 및 미술관 진흥법. 법률 제19592호). 하지만 기록관은 설립주체나 유형에 따라 법령상 명시된 구분 기준은 없다. 참고 가능한 부분은 국가기록원이 데이터를 개방하면서 기록관을 관할하는 기관의 성격에 따라 기록관의 목록을 구분한 것이다. 데이터에 따르면, 중앙행정기관 및 특별행정기관, 지방자치단체, 교육청 및 교육지원청, 공공기관, 대학으로 구분된다. 군기관의 기록관도 있지만 보안상의 이유로 따로 목록을 개방하지 않는다. 따라서 기록관에서 사용하는

〈표 2〉 기관별 메타데이터의 집계 현황

구분	전체 메타데이터 현황(개)
도서관	53
박물관	95
미술관	90
기록관	13

‘기관유형’ 항목은 공통 메타데이터에 포함되지는 않지만 의미상 도서관과 박물관에서 사용하는 ‘설립주체’와 유사하고 기록관을 분류할 수 있는 주요한 지표 중 하나라고 판단하여 추가적으로 추출하였다.

추가한 또 다른 메타데이터는 ‘기관종류’ 항목으로 기관의 종류를 나타내기 위해 사용하였다. 즉, 해당 기관이 도서관, 박물관, 미술관, 기록관 중 어떤 기관에 해당되는지 알려주는 역할을 한다. 각 기관의 데이터가 개별적으로 존재한다면 이 메타데이터는 큰 의미가 없지만, 여러 종류의 기관에 대한 데이터가 함께 있는 경우 구분해주는 메타데이터가 필요하다. 추출한 공통 메타데이터 항목은 〈표 3〉에 정리되어 있으며, 각 기관별로 특정 메타데이터 항목이 사용되었는지 확인할 수 있다.

추출한 공통 메타데이터를 사용하여 기관별 한 벌의 데이터를 생성하였다. 이 데이터는 지식그래프로 변환할 때 사용할 데이터로, 수집한 기관 선정용 데이터에 추출한 공통 메타데이터를 연결하여 데이터를 구축하였다. 만약 공통 메타데이터로 추출한 항목이 특정 기관의 데이터에서 사용되지 않은 경우, 데이터 정제 과정을 통해 채울 수 있는 인스턴스는 채우고자 하였다. 값을 채우지 못하더라도 해당 메타데이터를 추가는 하되, 빈 값으로 유지하였다.

〈표 3〉 공통 메타데이터와 기관 데이터별 사용 여부

메타데이터	도서관 포함 여부	박물관 포함 여부	미술관 포함 여부	기록관 포함 여부
시설명	O	O	O	O
시도	O	O	O	O
시군구	O	O	O	O
도로명주소	O	O	O	O
전화번호	O	O	O	O
홈페이지	O	O	O	O
개관연도	O	O	O	
부지면적	O	O	O	
소개		O	O	O
위도	O	O	O	O
경도	O	O	O	O
평일시작시각	O	O	O	
평일종료시각	O	O	O	
공휴일시작시각	O	O	O	
공휴일종료시각	O	O	O	
휴관정보	O	O	O	
우편번호	O	O		O
운영기관	O	O	O	
설립주체	O	O	O	
기관유형				O
기관종류	-	-	-	-

데이터 통합을 위해 기관의 유형에 구분없이 메타데이터 항목별로 동일한 인스턴스 값의 형식을 가져야 한다. 형식은 다수의 데이터에서 사용하는 형식을 따랐다. 〈표 4〉는 본 연구에서 사용한 인스턴스의 형식이다. 형식이 필요없는 항목은 표에 포함되어 있지 않다.

3. 지식그래프 구축

3.1 데이터 정제

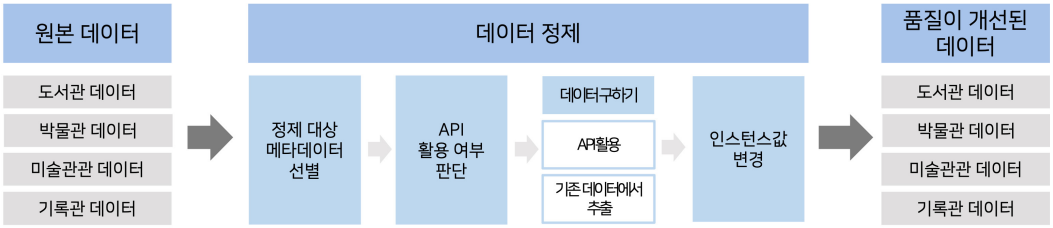
데이터를 지식그래프로 변환하기 전에 먼저 데이터를 정제하였다. 정제 대상 데이터는 지식그래프를 구축할 파일로, 각 기관별 한 벌의 데이터파일이며 총 네 개의 데이터파일이다. 각 파일은 선정된 21개의 공통 메타데이터 항목을 결합한 형태로 이루어져 있다. 정제를 먼저 진행하고 지식그래프로 변환하여 품질이 개선된 통합 지식그래프를 생성하고자 하였다. 정제과정은 〈그림 2〉를 참고한다.

데이터를 정제하는 목적은 크게 두 가지다. 첫 번째 목적은 기존 데이터의 품질을 개선하는 것이다. 본 연구의 목적이 고품질의 스몰데이터를 구축하는 것이므로, 원본 데이터보다는 품질이 개선된 테이블 데이터를 대상으로 지식그래프를 구축하고자 하였다. 따라서 데이터의 형식에 오류가 있다면 변경하고, 일부 인스턴스에 빈 값이 있다면 채우는 정제를 진행하였다.

두 번째 목적은 공통 메타데이터를 추출하고 메타데이터를 통합하는 과정에서 본래의 원본 데이터에는 없던 메타데이터가 추가되는 경우가 있다. 이러한 경우, 데이터가 아예 존재하지 않으므로 데이터 정제를 하면서 입력 가능한 데이터는 채우기 위해 진행하였다. 예를 들어

〈표 4〉 메타데이터별 인스턴스 형식

메타데이터	인스턴스 형식	메타데이터	인스턴스 형식
시도	행정구역 공식 명칭	평일시작시각	HH:MM
시군구	행정구역 공식 명칭	평일종료시각	HH:MM
전화번호	11자리 이내 문자	공휴일시작시각	HH:MM
위도	NN.NNNNNN ..	공휴일종료시각	HH:MM
경도	NNN.NNNNNN ..	개관연도	YYYY



〈그림 2〉 정제과정 도식화

‘우편번호’ 메타데이터는 본래 미술관데이터에는 존재하지 않는다. 하지만 기존의 도로명주소 데이터를 변환해서 우편번호 데이터를 구하여 최대한 결측 데이터의 비율을 줄이고자 하였다.

데이터 정제는 파이썬 코드를 통해 진행하였으며, 코드로 해결 가능한 범위 내에서 정제를 진행하였다. 데이터의 이용자 입장에서 해결 가능한 데이터 오류 사항은 크게 두 가지다. 첫 번째는 빈 인스턴스 값을 채우는 것으로, 참고 가능한 다른 메타데이터의 인스턴스 값을 활용하여 결측 데이터를 해결하였다. 예를 들어 ‘도로명주소’ 메타데이터에 결측이 있다면, ‘위도’와 ‘경도’ 값을 읽어와서 도로명주소 형식의 데이터로 변환한다. 또는 ‘시도’나 ‘시군구’ 메타데이터의 결측 인스턴스는 ‘도로명주소’ 메타데이터에서 행정구역 데이터를 추출해서 대체할 수 있다. 21개의 메타데이터 항목을 검토한 결과, 서로 값을 변환할 수 있는 메타데이터의 종류는 주소와 관련된 메타데이터로, ‘시도’, ‘시군구’, ‘도로명주소’, ‘위도’, ‘경도’가 있다.

두 번째 방법은 인스턴스 값의 형식을 통일하는 것이다. ‘시도’와 ‘시군구’ 메타데이터와 ‘도로명주소’ 메타데이터의 인스턴스 값에 공식 지역 명칭이 사용되지 않은 경우, 공식 명칭을 사용하도록 정제하였다. 이 외에도 ‘위도’와 ‘경도’ 인스턴스 값에 자릿수가 부족하여 정확한 위치를 식별할 수 없다면 도로명주소의 값을 변환하여 위도와 경도 정보를 다시 구하였다.

행정구역명에 공식 명칭을 입력하는 또 다른 이유는 데이터 조정을 성공적으로 진행하기 위한 것도 있다. 데이터 조정을 통해서 기존에 구축되어 있던 지식그래프를 연결하면 더 많은 정보를 확인할 수 있다. 이 때 데이터 조정을 진행하는 연결지점은 바로 URI로, 연결하고자 하는 그래프와 동일한 URI를 생성해야 한다. 보통 인스턴스 값을 기준으로 URI와 매칭하여 URI를 생성한다. 따라서 인스턴스 값이 정확해야 URI를 생성할 수 있고 데이터 조정을 성공적으로 수행할 수 있다. 본 연구는 기구축되어 있는 행정구역 지식그래프와 데이터 조정을 진행하여 각 기관이 위치한 시도와 시군구에 대한 정보까지 연결하고자 하였다. 시도와 시군구 메타데이터 항목에 공식 명칭이 아닌, 축약명이 입력되어 있는 경우 공식 명칭으로 변경하였다. 〈표 5〉는 메타데이터별로 진행하는 정제 내용을 정리한 자료다.

데이터 정제 과정에서 주소, 행정구역과 관련된 데이터의 변환은 기업 API를 사용한다.

〈표 5〉 정제 내용과 적용되는 메타데이터

컬럼	문제상황		해결방법
위도	결측	자릿수 부족	도로명주소 데이터 변환
경도			
시도		축약어 사용	도로명주소 데이터에서 추출
시군구			
도로명주소			위도와 경도 데이터 변환
우편번호		-	위도와 경도 데이터 변환

본 연구는 Kakao[5]에서 제공하는 REST API를 사용한다. 이 API는 여러 메소드를 지원하는데, 위도와 경도 데이터로 도로명주소를 구하는 메소드와, 반대로 도로명주소로 위도와 경도 데이터를 구하는 메소드를 주로 사용하였다. API는 파이썬언어로 사용이 가능하여, 데이터 정제는 파이썬을 사용해서 진행하였다. 〈표 6〉은 데이터 정제 진행 전과 후의 우수한 품질을 갖춘 인스턴 수를 비교한 표로, 전반적으로 품질이 개선된 모습을 알 수 있다. 전체적으로 결측이나 형식 오류의 인스턴스 수가 줄어들고, 품질이 준수한 인스턴스의 비율이 늘어났다.

3.2 지식모델의 설계

네 종류의 기관을 지식그래프로 구축하여 의미적으로 연계함으로써 통합적인 탐색이 가능하다. 중요한 것은 현재 존재하는 기관뿐만 아니라 미래에 새로 설립될 기관의 정보도 수용가능한 체계를 마련하는 것이다. 일회성이 아닌 지속적으로 사용가능한 스몰데이터를 유지하기

〈표 6〉 데이터 정제 전과 후의 오류 데이터의 변화

기관	메타데이터	기관 수	우수한 품질을 갖춘 인스턴스			
			정제 전		정제 후	
			인스턴스 수	인스턴스 비율(%)	인스턴스 수	인스턴스 비율(%)
도서관	위도	1,212	902	74.42	1,195	98.6
	경도		902	74.42	1,195	98.6
	도로명주소		1,137	93.81	1,197	98.76
	시도		0	0	1,210	99.83
	시군구		0	0	1,210	99.83
박물관	위도	913	911	99.78	913	100
	경도		911	99.78	913	100
	도로명주소		63	6.9	905	99.12
	시도		12	1.31	913	100
	시군구		905	99.12	906	99.23
기록관	위도	1,178	605	51.36	609	51.7
	경도		605	51.36	609	51.7
	도로명주소		496	42.11	536	45.5
	시도		549	46.6	589	50
	시군구		3	0.25	540	45.84
미술관	위도	285	249	87.37	249	87.37
	경도		249	87.37	249	87.37
	도로명주소		253	88.77	275	96.49
	시도		0	0	285	100
	시군구		0	0	285	100
	우편번호		0	0	219	76.84

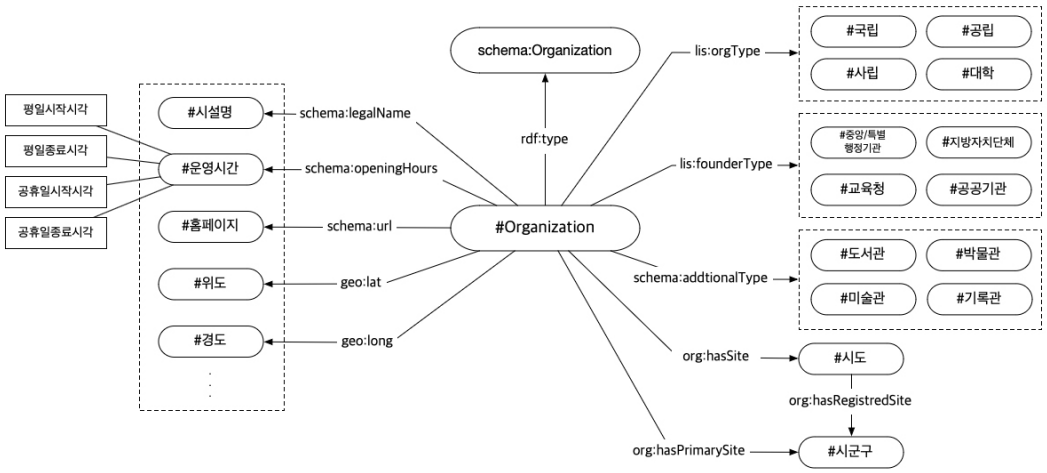
5) “Kakao API”, https://developers.kakao.com/, 2023년 11월 20일 접속

위한 기반이 필요하다. 이는 지식그래프로 변환하기 전, 지식모델을 설계함으로써 해결할 수 있다. 새롭게 추가되거나 확장되는 데이터가 일정한 RDF 기반의 지식모델을 준수하여 변환된다면 이질성없이 기존의 지식그래프와 연결될 수 있을 것이다. 따라서 3장에서 추출한 공통 메타데이터 21개 대상 속성을 지식모델을 설계하였다. 〈그림 3〉은 도서관, 박물관, 미술관, 기록관 중 한 곳의 기관이 21개의 속성과 연결된 구조를 표현한 것이다.

〈표 7〉은 세 개의 RDF 어휘를 사용하여 21개의 메타데이터에 속성을 매핑한 자료다. 사용한 RDF 어휘는 Organization Ontology, Schema.org, Basic Geo Vocabulary이며, 기존의 어휘로 표현하지 못 하는 메타데이터인 "설립주체"와 "기관유형"은 별도로 선언하여 사용하였다.

3.3 지식그래프의 변환

지식그래프에서 사용하는 URI 체계를 정의하였다. URI는 고유한 자원을 식별하기 위한 체계다. 필요한 URI 체계는 크게 두 개의 유형으로 나뉜다. 첫 번째 유형의 URI는 도서관, 박물관, 미술관, 기록관임을 의미하는 식별자 URI다. 각 기관을 의미하는 식별자를 조금씩 다르게 생성하여 네 종류의 기관에 대한 정보를 전체 통합한 지식그래프지만, 전체적인 정보 뿐만 아니라 도서관, 박물관, 미술관, 기록관 각 종류별 기관에 대한 데이터만 구분하여 확인할 수 있다. 예를 들어 도서관의 URI에 "library"를 축약한 "li"를 URI에 활용하여 "http://data.datahub.kr/lis-org/id/li/{일련번호}" 형식으로 URI를 생성하였다. 차례대로 박물관, 미술관, 기록관의 영문명을 축약한 "mu", "ag", "ac"를 입력하여 사용하였다. 두 번째 유형의 URI는 행정구역을 나타내기 위한 URI 체계로, "시도"와 "시군구" 메타데이터가 해당된다. 이 두 개의 메타데이터는 기존에 구축되었던 행정구역 지식그래프와 데이터 조정을 진행할 때 연결지점이 된다. 연결하고자 하는 지식그래프와 동



〈그림 3〉 설계한 지식 모델

〈표 7〉 공통 메타데이터별 RDF 어휘와 속성 매핑

메타데이터	어휘 종류	도메인	속성
시도	Organization Ontology	org:Organization	org:hasSite
시군구			org:hasPrimarySite
시설명	Schema.org	schema:Organization	schema:legalName
도로명주소			schema:address
전화번호			schema:telephone
개관년도			schema:foundingDate
부지면적			schema:areaServed
소개			schema:description
휴관정보			schema:description
우편번호			schema:location
운영기관			schema:parentOrganization
기관종류			schema:additionalType
평일시작시각		schema:LocalBusiness	schema:openingHours
평일종료시각			schema:openingHours
공휴일시작시각			schema:openingHours
공휴일종료시각			schema:openingHours
홈페이지		schema:Thing	schema:url
위도	Basic Geo Vocabulary	geo:Point	geo:lat
경도			geo:long
설립주체	(자체 생성)	lis:Organization	founderType
기관유형			orgType

일한 URI를 공유해야 데이터 조정이 가능하므로, 행정구역 지식그래프의 URI 체계를 준수하였다. 행정구역 지식그래프는 중앙대학교 HIKE 연구실에서 구축한 지식그래프다. 시도와 시군구, 읍면동 단위를 법정동코드를 이용해서 URI를 만들고, 행정구역별 인구와 관할하는 기관, 지역에 대한 설명 등을 속성으로 표현한 그래프다. 각 행정구역은 계층적으로 연결되어 있다. 행정구역 지식그래프의 URI는 "https://data.datahub.kr/administration/administrative-division/id/{법정동코드}" 형식을 사용한다.

또한 기관 사이의 연계를 위해 통제어휘를 정의해야 한다. 두 가지 유형의 통제어휘를 정의해야 하는데, 하나는 "기관종류"를 의미하는 메타데이터인 "lis:orgType"의 값으로 활용되는 통제어휘다. 이 통제어휘는 "http://data.datahub.kr/lis-org/category/{기관종류}" 체계를 따르며 "{기관종류}" 위치에 각 기관을 의미하는 영어단어가 사용되었다. 또 다른 통제어휘는 "설립주체"를 의미하는 메타데이터인 "lis:founderType"의 값으로 활용된다. 설립주체는 도서관, 박물관, 미술관이 법률상 같은 항목을 공유하는데, 추가적으로 박물관과 미술관만 '대학'을 사용한다. "http://data.datahub.kr/lis-org/library-type/{설립주체}" 형식을 따르며, "{설립주체}" 위치에 차례대로 국립, 공립, 사립, 대학을 의미하는 "national", "public", "private", "university"가 사용되었다.

파이썬 언어를 사용해서 테이블 형태의 데이터를 지식그래프로 변환하였다. 파이썬 라이브러리 중 RDFlib를 활용하여 그래프 데이터를 생성한다. RDFlib은 RDF 작업을 위한 패키지로, RDF/XML, 터틀(turtle) 등의 확장자로 직렬화를 지원한다.

터틀 형식의 그래프 데이터로 생성된 도서관, 박물관, 미술관, 기록관의 현황 데이터를 RDF 트리플 저장소인 GraphDB에 저장하였다. GraphDB는 OWL(Ontology Web Language) 기반의 그래프데이터베이스다. 〈표 8〉은 각 기관별로 생성된 진술문과 하나의 레포지토리에 세 개의 그래프 데이터 파일을 같이 저장한 후 생성된 진술문의 수를 정리한 자료다. 명시적 진술문과 추론된 진술문, 그리고 총 진술문의 수를 확인할 수 있다. 추론은 기존의 사실과 규칙을 기반으로 새로운 결론을 이끌어내는 것을 말한다(Chen, Jia, & Xiang, 2020). 테이블 형태의 데이터와 비교했을 때 갖는 큰 장점 중 하나다. 테이블 형태의 데이터는 엔티티간의 관계가 명시적으로 정의되었을 경우에만 관계를 확인할 수 있지만, 지식그래프는 기존에 정의된 관계를 바탕으로 엔티티간의 숨겨진 연결관계를 발견할 수 있다. 이는 기존의 데이터를 더욱 풍부하게 해준다.

3.4 검증과 시각화

GraphDB는 SPARQL 질의 기능을 지원한다. 통합 데이터를 대상으로 SPARQL 질의를 해서 기관의 종류에 상관없는 전체 탐색이 가능한지 검증하는 단계다. 기관의 "설립주체"가 "공립"인 도서관, 박물관, 미술관의 연결관계를 탐색하기 위한 질의를 입력하였으며, 실제 SPARQL 질의 내용은 〈표 9〉와 같다.

도서관, 박물관, 미술관은 도서관법과 박물관 및 미술관 진흥법에 의거하여 설립주체가 지방자치단체거나 교육청인 곳을 "공립"으로

〈표 8〉 트리플 저장소에 지식그래프를 저장한 후 생성된 진술문의 수

구분	명시적 진술문(개)	추론된 진술문(개)	총 진술문(개)
도서관	20,939	106	21,045
박물관	17,604	108	17,712
기록관	12,173	98	12,271
미술관	5,222	108	5,330
합계	54,860	108	54,968

〈표 9〉 SPARQL 질의문

```
select distinct ?category (count(distinct ?org) as ?cntOrg)
where { ?org a schema:Organization ;
        lis:founderType ?founder ;
        lis:orgType ?category.
  {FILTER regex(str(?founder), "public$", "i")}} groupby ?category
```

공통적으로 분류하고 있다. 하지만 기록관을 구분하는 공식적인 기준이나 관련 법령이 존재하지 않는다. 참고 가능한 사항은 국가기록원에서 기록관의 목록을 개방할 때 기관의 유형에 따라 구분하여 개방하지만, 공식적으로 사용되는 기준이라고 볼 수 없다. 따라서 본 쿼리의 결과로는 기록관의 검색이 불가능하다.

공립 도서관, 박물관, 미술관을 검색할 때 설립주체가 "공립"을 의미하는 URI인 "http://data.datahub.kr/lis-org/type/public"을 사용하는 기관을 검색하면 통합 검색이 가능하지만, 다른 URI를 사용하는 기록관은 해당 URI로 연결될 수 없다. 여러 종류의 기관에 대한 현황 정보를 통합하기 위해서는 사전에 일정한 통합규칙이나 통제어휘 사용 여부를 합의해야 한다. 또한 이 합의는 데이터 구축, 개방, 활용의 기반이 되는 정책적 체계가 마련되어야 진정한 통합이 가능하다. 기록관의 경우, 다른 기관과 다르게 설립주체를 기준으로 공식적으로 마련한 분류체계가 없다. 이는 정책상으로 분류체계가 먼저 마련되어야 하는 부분이다.

질의 결과는 〈표 10〉이다. 지식그래프로 변환된 기관 데이터 중, 도서관은 1,181곳, 박물관은 386곳, 미술관은 79곳이 공립에 해당된다.

지식그래프로 구축된 데이터를 대상으로 직접 SPARQL 쿼리를 이용하여 통합적인 탐색이 가능한지 실험을 진행하였다. 그 결과, 동일한 URI를 사용하거나 메타데이터 항목을 사용한다면 기관의 종류에 제약없이 기관 데이터를 탐색할 수 있었다. 지식그래프는 이기종의 정보를 의미적으로 연계하기 위한 효과적인 방법 중 하나이며, 관리와 검색 측면에서 활용성을 증진시킬 수 있는 방안이다.

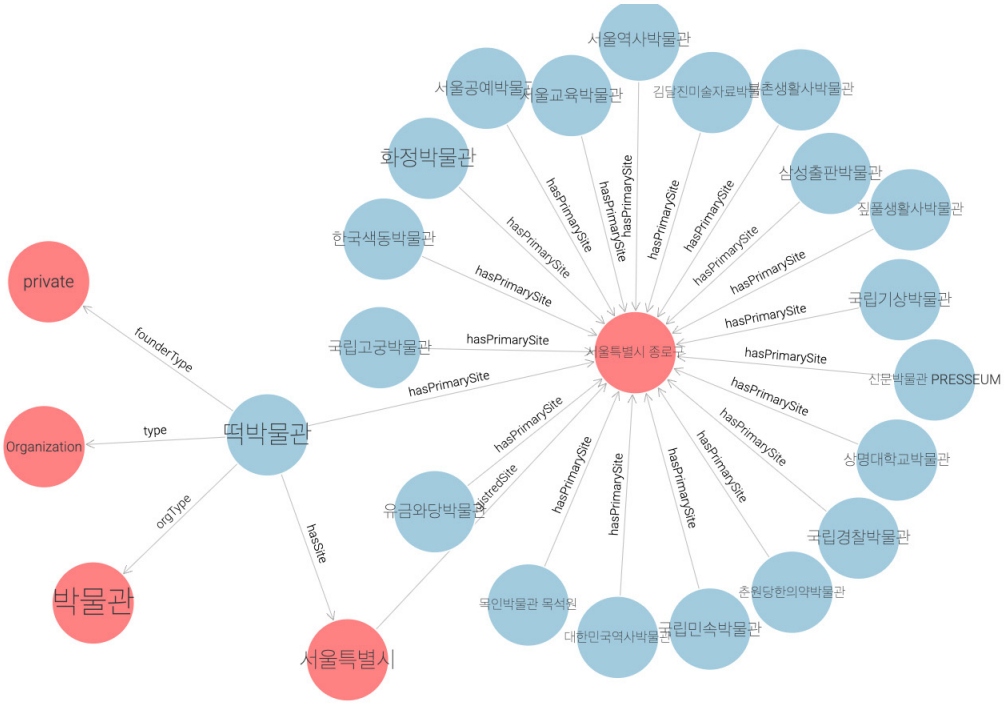
〈그림 4〉는 "서울특별시 종로구" 행정구역에 위치한 여러 문화예술 기관이 연결된 모습을 시각화한 자료다. "서울특별시 종로구"를 의미하는 URI를 기준으로 여러 종류의 기관이 연결되어 있다. 통합된 데이터를 시각화를 이용해 서비스하는 것은 지식그래프 개념에 대해 생소한 사람도 쉽게 구조를 파악할 수 있도록 도움을 줄 수 있으므로, 지식그래프 개념이 생소한 이용자에게 유용하게 사용될 수 있는 기능 중 하나임을 확인하였다.

4. 결 론

본 연구는 웹에 분산적으로 개방되고 있는 문화예술기관의 기본정보를 지식그래프 형태로 통합하여 스몰데이터를 구축하였다. 법률적 기준에 근거하여 지식그래프로 통합할 기관인 도서관 1,212곳, 박물관 913곳, 미술관 285곳과 기록관 1,304곳을 정의하였으며, 데이터포탈이나 웹 크롤링을 통해 각 기관의 현황데이터를

〈표 10〉 SPARQL 질의문 결과

기관 종류(category)	공립인 곳의 수량(cntOrg)
"http://data.datahub.kr/lis-org/category/library"	1,181
"http://data.datahub.kr/lis-org/category/museum"	386
"http://data.datahub.kr/lis-org/category/art_gallery"	79



〈그림 4〉 "서울특별시 종로구"를 기준으로 연결된 문화예술 관련 기관

수집하고 메타데이터 항목을 분석하였다. 그 결과 19개의 공통 메타데이터를 추출하였으며, 기관의 통합을 위해 필요하다고 판단되는 메타데이터 2개를 추가적으로 추출하여 총 21개의 메타데이터를 선정하였다. 선정한 메타데이터를 중심으로 각 기관이 연결된 지식그래프를 구축하여 문화예술기관을 통합적으로 탐색할 수 있는 스몰데이터를 구축하였다. 지식그래프로 변환하기 전, 고품질의 스몰데이터를 구축하기 위해 데이터 정제 작업을 진행하였다. 메타데이터 항목을 추가하면서 생기는 결측 데이터를 해결하고 인스턴스 형식을 통일하기 위해 진행하였으며, 외부 API를 활용하여 파이썬으로 자동화된 정제 코드를 작성하여 적용하였다. 그 후 선정된 21개의 메타데이터 항목을 대상으로 지식모델을 설계하였는데, 각 항목별로 적절한 RDF 어휘와 속성을 부여하였고 기존 어휘로 표현하지 못 하는 2개의 메타데이터 항목은 신규로 정의한 속성을 사용하였다. 실제 지식그래프의 변환은 파이썬 라이브러리인 RDFLib을 활용하였다. 변환된 데이터는 그래프데이터베이스인 GraphDB에 저장하고 SPARQL 질의를 통해 검증하였다.

지식그래프는 다른 데이터와의 확장이 가능한 형태이며 복잡한 관계도 직관적으로 표현할 수 있다는 장점이 있다. 또한 특정 행정구역이나 기관의 종류라는 특성은 여러 종류의 기관을 연결할 수 있는 접근점이 되고, 이 접근점을 기준으로 기관의 탐색이 가능하다. 지식모델을 먼저 설계하고 이를 기반으로 지식그래프를 구

현하였다. 이는 새롭게 설립되는 기관이 있더라도 일정한 설계규칙을 준수하여 데이터를 변환한다면 기존의 지식그래프에 연결하여 데이터의 확장이 가능함을 의미한다. 즉, 본 연구를 통해 지속적인 관리와 활용이 가능한 스몰데이터를 구축하였다.

기관의 특성상 사용하는 메타데이터 항목은 다양할 수 있다. 예를 들어 도서관의 기관정보는 소장자료를 도서자료, 비도서자료, 연속간행물 등으로 세부적으로 나눠서 개방하고 있다. 본 연구는 공통 메타데이터 항목 21개에 한정하여 지식그래프를 구현하였다. 향후 특정 종류의 기관에서 고유하게 사용하는 메타데이터를 분석하여 도메인 특화 메타데이터를 추가한다면, 문화예술기관의 통합적인 탐색과 함께 다른 데이터를 참고하지 않더라도 각 기관의 세부적인 정보까지 탐색할 수 있는 데이터를 구축할 수 있을 것이다.

데이터의 진정한 연계를 위해 정책적 기반도 마련되어야 한다. 예를 들어, 도서관, 박물관, 미술관은 각 관할법에 의거하여 기관을 설립한 주체의 유형에 따라 기관을 구분한다. 도서관은 국립, 공립, 사립으로 구분하고 박물관과 미술관은 국립, 공립, 사립, 대학으로 구분하지만 기록관은 공식적인 유형이 없다. 국가기록원에서 기록관 목록을 개방할 때 기록관을 운영하는 기관의 유형에 따라 데이터를 구분하여 개방하지만, 법률적인 기준은 존재하지 않는다. 임의적으로 기록관의 유형을 나누어 도서관, 박물관, 미술관과 같은 체계를 유지하는 것은 자의적인 판단이 될 수 있다. 따라서 정책적으로 기록관을 구분하는 근거가 마련된다면 데이터 통합과 활용에 긍정적인 변화를 가져올 수 있을 것이다.

참 고 문 헌

김학래 (2017). 지식그래프. 서울: 커뮤니케이션북스.

도서관법. 법률 제19592호.

박물관 및 미술관 진흥법. 법률 제19592호.

복경수, 유재수 (2017). 4차 산업혁명에서 빅데이터. 정보과학회지, 35(6), 29-39.

윤소영 (2013). 공공데이터 활용을 위한 링크드 데이터 국가 연계체계 구축에 관한 연구. 정보관리학회지, 30(1), 259-284. https://doi.org/10.3743/KOSIM.2013.30.1.259

이용주 (2014). 링크드 데이터 구축 및 검색 기법. 한국정보처리학회 학술대회논문집, 21(2), 1057-1060.

Abu-Salih, B. (2021). Domain-specific knowledge graphs: a survey. Journal of Network and Computer Applications, 185, 103076. https://doi.org/10.1016/j.jnca.2021.103076

Berners-Lee, T. (1998). Semantic Web Road Map. https://www.w3.org/DesignIssues/Semantic.html

Bush, V. (1945). As we may think. The Atlantic Monthly, 176(1), 101-108. https://doi.org/10.1145/227181.227186

Chen, X., Jia, S., & Xiang, Y. (2020). A review: knowledge reasoning over knowledge graph. Expert Systems with Applications, 141, 112948. https://doi.org/10.1016/j.eswa.2019.112948

Daruna, A., Das, D., & Chernova, S. (2022). Explainable Knowledge Graph Embedding: Inference Reconciliation for Knowledge Inferences Supporting Robot Actions. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1008-1015). https://doi.org/10.48550/arXiv.2205.01836

Fensel, D., Simsek, U., Angele, K., Huaman, E., Karle, E., Panasiuk, O., Toma, L., Umbrich, J., & Wahler, A. (2020). Introduction: What Is a Knowledge Graph?. Knowledge Graphs. https://doi.org/10.1007/978-3-030-37439-6_1

Grau, B. C., Horrocks, I., Kazakov, Y., & Sattler, U. (2008). Modular reuse of ontologies: theory and practice. Journal of Artificial Intelligence Research, 31, 273-318. https://doi.org/10.1613/jair.2375

Gutiérrez, C. & Sequeda, J. F. (2021). Knowledge graphs. Communications of the Association for Computing Machinery, 64(3), 96-104. https://doi.org/10.1145/3418294

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra G., Navigli, R., Neumaier, S., Ngonga A., Polleres, A., Rashid, S., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. Association for Computing Machinery Computing Surveys (Csur), 54(4), 1-37. https://doi.org/10.1145/3447772

Iliadis, A., Acker, A., Stevens, W., & Kavakli, S. B. (2022). One Schema to Rule Them All: How Schema.org Models the World of Search. Journal of the Association for Information Science and Technology. https://doi.org/10.1002/asi.24744

Kollar, T., Berry, D., Stuart, L. M., Owczarzak, K., Chung, T., Mathias, L., Kayser, M., Snow, B., & Matsoukas, S. (2018). The alexa meaning representation language. North American Chapter of the Association for Computational Linguistics, 177-184. https://doi.org/10.18653/v1/N18-3022

Latzko-Toth, G., Bonneau, C., & Millette, M. (2017). Small data, thick data: thickening strategies for trace-based social media research. The SAGE Handbook of Social Media Research Methods, 199-214. https://doi.org/10.4135/9781473983847

Mandal, S. (2022). Integration of linked open data authorities with open refine: a methodology for libraries. Library Philosophy & Practice, 1-9.

Miles, A. & Pérez-Agüera, J. R. (2007). Skos: simple knowledge organisation for the web. Cataloging & Classification Quarterly, 43(3-4), 69-83. https://doi.org/10.1300/J104v43n03_04

Mongiovì, M., Recupero, D. R., Gangemi, A., Presutti, V., Nuzzolese, A. G., & Consoli, S. (2015). Semantic reconciliation of knowledge extracted from text through a novel machine reader. In Proceedings of the 8th International Conference on Knowledge Capture, 1-4. https://doi.org/10.1145/2815833.2816945

Monnin, P., Raïssi, C., Napoli, A., & Coulet, A. (2019). Knowledge Reconciliation with Graph Convolutional Networks: Preliminary Results. In DL4KG2019-Workshop on Deep Learning for Knowledge Graphs.

Sanchez Alonso, S. & Garcia Barriocanal, E. (2006). Making use of upper ontologies to foster interoperability between SKOS concept schemes. Online Information Review, 30(3), 263-277. https://doi.org/10.1108/14684520610675799

Sohmen, L. & Rossenova, L. (2022). Open refine to wikibase: a new data upload pipeline. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, 1-2. https://doi.org/10.1145/3529372.3530919

Strickland, E. (2022). Andrew Ng, AI minimalist: the machine-learning pioneer says small is the new big. g. IEEE Spectrum, 59(4), 22-50. https://doi.org/10.1109/MSPEC.2022.9754503.

Tiwari, S., Al-Aswadi, F. N., & Gaurav, D. (2021). Recent trends in knowledge graphs: theory and practice. Soft Computing, 25, 8337-8355. https://doi.org/10.1007/s00500-021-05756-8

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: a survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743. https://doi.org/10.48550/arXiv.2107.07842

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Bok, Kyoung-soo & Yoo, Jae-soo (2017). Big data in the fourth industrial revolution. Journal of the Korean Data And Information Science Sociaty, 35(6), 29-39.

Kim, Hak-lae (2017). Knowledge Graph. Seoul: Communicationbooks.

Lee, Yong-ju (2014). Building and retrieval techniques of linked data. Korea Information Processing Society, 21(2), 1057-1060.

Libraries Act. Act No. 19592.

Museum And Art Gallery Support Act. Act No. 19592.

Yoon, So-Young (2013). A study on national linking system implementation based on linked data for public data. Journal of the Korean Society for Information Management, 30(1), 259-284. https://doi.org/10.3743/KOSIM.2013.30.1.259