

공공데이터포털에 공개된 지방자치단체의 공공데이터 활용 방안

김 동 준¹ · 김 현 지¹ · 송 채 은¹ · 양 지 우¹ · 김 학 래^{2*}

¹중앙대학교 문헌정보학과 학사과정

^{2*}중앙대학교 문헌정보학과 교수

Methods for Utilising Local Government's Public Data Released to The Public Data Portal

Dongjun Kim¹ · Hyeonji Kim¹ · Chaeun Song¹ · Jiwoo Yang¹ · Haklae Kim^{2*}

¹Bachelor'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]

대한민국 정부는 지속적으로 공공데이터를 개방하고 있고, 산업적 활용을 위해 다양한 정책을 추진하고 있다. 정부가 공공데이터를 지속적으로 개방하고 있음에도, 데이터 활용은 여전히 제한적이다. 공공데이터는 낮은 수준의 데이터 품질, 이종의 데이터를 연계하기 위한 표준 등 다양한 이슈를 갖고 있다. 한편, 데이터 수요를 적시에 파악하고 개방하기 위한 프로세스, 개방된 공공데이터의 활용 현황의 진단 등 공공 데이터 환경에 대한 포괄적 진단이 필수적이다. 본 논문은 공공데이터포털에 개방된 지방자치단체의 데이터세트의 활용도를 평가한다. 데이터 활용도는 개방된 데이터를 사용자의 조회와 다운로드한 수치로 측정한다. 공공데이터가 지속적으로 공개되고 있지만, 개방된 데이터가 민간 또는 기업에 의해 실제 활용되지 않고 있다. 본 논문은 지방자치단체의 공공데이터의 활용 현황을 진단하고, 공공데이터 활용 환경을 개선하기 위한 방안을 제안한다.

[Abstract]

The Korean Government is continuously disclosing public data and is pursuing various policies for industrial use. Although the government continues to public data, the use of public data in the private sector is still limited. There are some issues associated with public data, such as improvement of low quality and standards for connecting between heterogenous dataset. Also, comprehensive diagnosis is essential, including a process of identifying and releasing on-demand datasets in a timely manner and evaluation for public data that has been released. This paper evaluates the utilisation of local governments' public data released to the public data portal. The data utilisation is measured by the user's views and downloads of individual datasets. Although public data is constantly being disclosed, they are not actually being used by private sectors or enterprises. This paper diagnoses the current state of use of public data by local governments and proposes a means to improve the environment for using public data.

색인어 : 데이터포털, 정부 데이터, 데이터 활용성, 데이터 분석, 지방자치단체

Key word : data analysis, Local government, local government, public data, government data, usability, etc.

<http://dx.doi.org/10.9728/dcs.2021.22.3.445>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 14 December 2020; **Revised** 20 January 2021

Accepted 20 January 2021

***Corresponding Author; Haklae Kim**

Tel: +82-2-820-5561

E-mail: haklaekim@cau.ac.kr

I. 서론

대한민국 정부는 공공데이터를 지속적으로 개방하고 있다. '공공데이터의 제공 및 이용활성화에 관한 법률'에 따르면, 공공데이터는 공공기관이 만들어내는 모든 자료나 정보, 국민 모두의 소통과 협력을 이끌어내는 공적인 정보이다 [1]. 공공데이터 정책은 정부의 투명성을 확보하는 것을 목표로 시작했으나, 데이터 중심 경제를 실현하는 핵심 자원으로 인식되고 있다. 공공기관이 관리하는 데이터는 다양한 경제·산업적 요구가 반영되어 있다. 지도, 날씨, 교통 정보는 국민의 일상생활에 밀접하고, 한국인의 얼굴표정, 목소리, 필기체 정보는 인공지능 분야에서 활용도가 높은 공공데이터이다.

그러나, 공공데이터의 효과적 활용은 제한적이다. 첫째, 데이터 개방 절차가 복잡하고 일관성 있는 가이드라인이 미흡하다. 공공데이터 관리지침은 공공데이터의 생성에서 사후관리에 이르는 단계별 기준을 제공하고 있지만, 개별 기관이 데이터를 공개하는데 필요한 세부 규칙을 상세하게 표현하지 않는다 [2]. 비록 수요가 확인된 데이터가 존재해도 데이터의 개방은 개인정보보호, 개별 기관의 특수한 상황에 따라 제한될 수 있다. 둘째, 개방된 데이터는 정확성, 완전성을 포함하는 품질과 활용 측면의 문제를 갖고 있다 [3]. 양질의 데이터는 부가적인 작업 없이 활용할 수 있기 때문에, 데이터 활용을 위한 필수조건이다. 공공데이터의 품질 이슈는 다양하게 논의하고 있지만, 데이터 거버넌스를 고려한 정책은 충분히 논의되지 않는 것이 현실이다. 현재 대부분의 대응은 메타데이터 수준에 머물러 있다. 셋째, 공공데이터의 활용 현황을 정량적으로 논의하는 데에 한계가 있다. 공공데이터포털은 조회수를 누적한 기준으로 '인기 데이터'를 제공하고, 데이터 활용과 시각화 사례를 별도로 기능으로 제공하고 있다. 이 기능은 관심 또는 인기 있는 데이터를 파악하는 데 유용할 수 있지만, 데이터가 실제 활용되었는지 파악하는 데 한계가 있다. 특히, 지방자치단체의 공공데이터 활용은 상세히 분석되지 못하고 있다 [4]-[5]. 해외의 경우, 중앙정부와 더불어 지방정부가 운영하는 데이터 포털이 활성화되었고, 이에 대한 활용성 분석도 활발히 진행되고 있다 [6].

본 연구는 지방자치단체가 공공데이터포털에 개방한 데이터셋의 활용도를 분석하고, 데이터 활용을 활성화시키기 위한 방안을 제안한다. 먼저 공공데이터포털에서 제공하는 데이터셋을 지방자치단체별로 분류하고, 메타데이터에서 추출한 정보를 바탕으로 지방자치단체별로 데이터 현황과 활용도를 분석한다. 공공데이터 활용도는 특정 데이터셋에 대한 조회수와 다운로드수의 관계로 측정한다.

본 논문의 구성은 다음과 같다. 2장은 공공데이터의 이용 활성화를 위한 관련 연구를 소개한다. 3장은 데이터 활용성을 측정하기 위한 연구 방법과 지표를 설명한다. 4장은 데이터 분석 결과를 정리하고, 5장에서 활용성을 개선하기 위한 방향을 논의한다. 마지막으로 6장은 연구 결과를 요약하고 향후 연구 방향을 소개한다.

II. 관련 연구

공공데이터는 다양한 데이터를 자유롭게 활용하고 재배포할 수 있는 특징이 있다. 공공데이터는 공공기관과 지방자치단체를 통해 지속적으로 공개되고 있다 [7]. 지방자치단체의 개방 현황과 주제 분야에 따른 데이터세트의 제공은 지방자치단체별, 기관별, 분야별로 관점에 따라 다양하게 연구되고 있다 [8].

공공데이터 활용은 활성화 정책, 평가지표, 데이터 품질 등 다양한 주제로 연구되고 있다. [7]은 공공데이터포털의 지방자치단체/기관에 따라 데이터 조회수, 다운로드수, 데이터수정일, 파일 형식을 분석하고, 공공데이터 활성화에 대한 개선사항을 제안하고 있다. [9]은 공공데이터의 제공기관인 중앙정부와 지방자치단체 사이의 공공데이터 정책과 현황을 분석하고 있다. 이에 따르면, 지방자치단체는 행정적, 재정적 상황이 다르기 때문에 일관성 있는 공공데이터 정책 추진을 위해 포괄적인 가이드라인의 제공이 필요하다고 주장한다.

그러나, 공공데이터 규모의 폭발적 확대는 데이터의 탐색과 활용에 대한 다양한 이슈를 만들고 있다. 홍연웅 [10]은 데이터 품질의 중요성과 표준화를 강조하고, 정부 기관의 데이터 공유 활성화를 위해 공공데이터와 빅데이터의 통합 지원을 제안하고 있다. 공공데이터를 링크드 데이터로 표현하고 의미적으로 질의할 수 있는 방안도 활발하게 연구되고 있다. 특히, 오픈 데이터 플랫폼에서 웹 표준을 적용하는 방안에 대한 연구가 유럽 국가를 중심으로 진행되고 있다 [11]. Machova et. al.은 국가별 오픈데이터포털의 활용성을 데이터의 탐색가능성, 접근성, 재사용성을 기준으로 평가하고, 활용성을 개선하기 위한 가이드라인을 제안하고 있다 [12].

데이터의 품질은 데이터 활용을 위한 핵심적인 요소이다. 공공데이터의 품질관리를 이론적인 모델 [13], 품질 관리 프로세스와 수준 평가 [14]를 위한 연구가 진행되고 있다. 김학래 [15]는 개방표준 데이터의 품질을 완전성과 정확성으로 진단하고, 공공데이터 관리지침의 개선사항과 정부 정책의 개선 방향을 제안한다. 공공데이터는 다양한 데이터세트의 결합 또는 연계를 통해 활용성을 높일 수 있다. 김학래 [3]는 데이터세트의 컬럼명을 표준화하고, 데이터세트 사이의 공통 관계를 개념적으로 분석하여 데이터의 연관 관계를 탐색하는 방안을 제안하고 있다. 데이터포털의 목록은 W3C의 권고안으로 발표된 DCAT [16]을 적용하는 방안이 활발히 논의되고 있고, 개별 국가의 상황에 따라 어휘를 확장하는 사례가 보고되고 있다 [16]-[20].

선행연구는 데이터의 표준과 품질 [21]-[22], 법률·제도적 관점 [23]에서 개선의 필요성을 지적하고 있는 반면, 본 연구는 본 연구는 공공데이터포털에 공개된 데이터세트의 활용을 사용자의 활동을 기반으로 측정한다. 더불어, 지방자치단체에서 개방한 데이터세트의 활용여부와 업데이트 여부를 평가함으로써 향후 개선을 위한 구체적인 근거를 제시할 수 있다.

III. 연구 방법

3-1 데이터 수집

데이터 수집은 공공데이터포털¹⁾에 제공된 지방자치단체의 데이터셋을 대상으로 한다. 공공데이터포털은 지방자치단체를 분류하는 기능을 제공하지 않고 있기 때문에 지방자치단체명으로 검색한다. 검색결과는 파이썬의 셀레니엄 (Selenium)과 뷰티풀 수프 (Beautifulsoup) 라이브러리를 이용해 개별 데이터셋의 메타데이터와 데이터 파일을 함께 수집한다. 2020년 08월 06일 기준으로 수집한 파일데이터 총 수량은 28,194건이며, 이 중 지방자치단체명으로 수집한 데이터는 16,057건이다. 수집한 데이터셋은 13,901건으로 2,156건의 데이터는 검색어와 관련 없는 데이터이다. 공공데이터포털의 규모로 보면, 지방자치단체의 파일데이터는 약 49%를 차지한다. 공공데이터포털에서 제공하는 메타데이터 요소는 다음과 같다.

Page_URL, 제공기관유형, 분류체계, 상세분류체계, 제공기관, 확장자, 파일이름, 파일데이터명, 조회수, 다운로드수, 주기성 데이터수, 등록일, 수정일, 관리부서명, 보유 근거, 수집방법, 업데이트 주기, 차기 등록 예정일, 매체유형, 전체 행, 데이터한계, 키워드, 제공형태, URL, 설명, 기타 유의사항, 비용부과유무, 비용부과기준 및 단위, 이용허락범위

3-2 데이터 활용 지수

공공데이터포털의 사용자는 필요한 데이터를 탐색하고, 데이터의 내용을 확인하고 다운로드해서 활용한다. 사용자가 데이터를 획득하는 기준을 측정하기 위한 목적으로 사용자의 데이터 탐색 단계를 다음과 같이 가정한다. 첫째, 데이터 사용자는 공공데이터포털의 검색 기능에 특정한 키워드를 제공한다. 둘째, 검색 결과에서 특정한 데이터셋을 선택하고, 메타데이터의 확인을 통해 다운로드 여부를 결정한다. 셋째, 다운로드를 하지 않을 경우, 재탐색을 수행한다. 결과적으로, 사용자의 데이터 활용은 조회와 다운로드의 관계로 정의할 수 있다. 수식 1은 특정 데이터셋 A의 데이터 활용지수(DPVA)를 측정하기 위한 계산식이다. 특정 데이터 A의 활용지수는 전체 다운로드 수를 전체 조회수로 나눈 값이다.

$$DPV_A = \left(\frac{\text{전체 다운로드수}_A}{\text{전체 조회수}_A} \right) \quad (1)$$

높은 데이터 활용지수는 조회한 데이터셋을 다운로드하는 빈도가 높다는 의미이다. 특정한 데이터셋의 조회수가 높을 수 있지만, 사용자가 데이터셋을 다운로드하지 않는다면 데이터의 직접적인 활용이 이루어지지 않는다고 가정한다.

1) <http://data.go.kr>

3-3 데이터의 갱신 주기

행정안전부의 공공데이터 관리지침에 관한 고시에서는 위와 같은 필요성에 대해 공공기관은 공공데이터포털에 제공하는 공공데이터의 최신성, 정확성을 유지할 수 있도록 관리해야 하며, 제공데이터의 업데이트 예정일 도래 시, 예정일에 맞춰 데이터 최신화가 이뤄질 수 있도록 관리하고 예정일 이전에 반드시 현행화하도록 조항으로써 명시해놓고 있다 [2]

공공데이터포털에 공개된 데이터는 제공기관이 지속적으로 관리해야 되는 의무가 있다. 개별 데이터셋의 갱신 여부는 ‘수정일’, ‘등록일’, ‘차기 등록예정일’, ‘업데이트 주기’ 항목을 이용해 검증할 수 있다. ‘차기 등록 예정일’ 컬럼은 데이터셋의 새로운 버전을 등록할 시점을 날짜로 표시하고 있고, ‘업데이트 주기’ 컬럼은 데이터의 갱신 주기를 표현한다. 업데이트 주기는 ‘연간’, ‘반기’, ‘분기’, ‘월간’, ‘주간’, ‘일간’ 등 범주형 데이터의 특성을 갖는다. 갱신 여부는 다음의 규칙을 적용한다.

- 데이터 갱신 미준수율: 개별 데이터 세트의 수집일자 (2020년 8월 6일)와 ‘차기 등록 예정일’을 비교하여, 데이터의 갱신 여부를 확인한다.
- 갱신 준수율: 개별 데이터셋이 ‘업데이트 주기’에 맞게 갱신되었는지 여부를 확인한다.

‘업데이트 주기’의 컬럼값에서 ‘연간’, ‘주간’, ‘일간’은 각각 365일, 7일, 1일의 수치로 치환하고, ‘반기’, ‘분기’, ‘월간’은 1년을 각각 2, 4, 12로 나눈 수치인 182.5일, 91.25일, 30.42일로 치환한다. 다만, ‘업데이트 주기’의 값이 공백이거나 ‘수시’로 입력된 것은 분석에서 제외한다. 갱신 준수 여부는 ‘수정일’부터 데이터 수집일자의 간격을 기준으로 하고, 수치로 치환된 업데이트 주기가 수정 경과일보다 크면 ‘준수’, 작으면 ‘미준수’로 구분한다.

IV. 분석 결과

4-1 데이터 현황

공공데이터포털에 제공하는 지방자치단체의 데이터는 총 13,901건이다. 표 1에서 보듯이, 서울특별시 2,631건으로 가장 많은 데이터를 제공하고 있고, 부산광역시와 경기도가 각각 1,652건, 1,444건을 제공하고 있다. 그러나 3개 지방자치단체를 제외하면 1,000개 이상의 데이터셋을 제공하는 기관이 없고, 세종특별자치시는 66개로 가장 적다. 공공데이터포털에서 정의한 분류체계를 기준으로 보면, ‘공공행정’ 2,735건(19.67%)으로 비중이 가장 높고, 산업고용(1,624건)과 문화관광(1,523건)도 높은 비율을 차지한다. 통일외교안보와 법률이 각각 27건, 3건인 상황에서, 지방자치단체가 제공하는 데이터는 특정 주제에 집중된 경향을 보인다.

표 1. 지방자치단체별 데이터셋과 이상치 규모

Table 1. The number of datasets and outliers by local governments

Local Government	Seoul	Gyeonggi	Incheon	Chungbuk	Chungnam	Sejong	Daejeon	Jeonbuk	Jeonnam	Gwangju	Jeju	Gyeongbuk	Gyeongnam	Busan	Ulsan	Daegu	Gangwon
Whole data	2631	1444	957	583	653	66	337	994	757	724	361	705	768	1652	341	562	366
Missing values	259	173	143	67	69	4	149	129	93	91	31	102	100	132	43	75	40

4-2 데이터 활용 (DPV)

모든 지방자치단체의 DPV의 평균은 0.3이고, 중앙값과 표준편차는 각각 0.219와 0.260이다. 분석과정에서 DPV의 이상치는 제외시켰다. 지방자치단체별 DPV의 분포는 그림 1과 같다. 중앙값의 비대칭적 분포는 조회와 다운로드의 관계가 지방자치단체의 데이터에 따라 다르다는 것을 보여준다. 강원도(0.371), 경기도(0.345), 경상남도(0.322) 등 11개 시도는 전체 중앙값보다 높다. 반면, 광주광역시(0.166), 서울특별시(0.127), 제주특별자치도(0.107) 등 6개 지방자치단체는 전국의 DPV 중위값보다 낮다. 전라북도, 부산광역시, 울산광역시, 광주광역시, 서울특별시, 제주특별자치도의 데이터 규모는 이상치를 제외하면 6,018 건이고, 전체 데이터의 49.32%에 해당한다. 따라서, 6개 지방자치단체는 데이터 활용도를 개선하기 위한 노력이 필요하다.

그렇다면, 제공되는 데이터 규모는 DPV에 긍정적인 영향을 줄 수 있을까? DPV의 분포로 보면, 데이터 규모는 사용자의 데이터 활용에 크게 영향을 미치지 않는다. 예를 들어, 서울특별시, 부산광역시는 데이터의 규모로 보면 상위 순위이고, 대전광역시와 세종특별자치시는 낮은 순위이다. 반면 DPV의 평균값을 기준으로 서울특별시와 부산광역시는 전체 지방자치단체에서 각각 13번째, 15번째에 위치하고, 대전광역시와 세종특별자치시는 그보다 상위에 위치한다. 또한 제주특별자치도와 울산광역시는 제공하는 데이터 규모가 낮고, DPV 기준 순위도 낮게 나타난다. 경기도는 데이터 규모로 3위, DPV 평균값 기준으로 2위이고, 데이터 규모가 낮은 강원도는 높은 DPV 순위를 갖는다.

요약해 보면, 데이터의 제공 규모는 사용자의 데이터 활용에 직접적인 영향을 미치지 않는다. 그러므로 공공데이터의 활용성 제고는 데이터의 양적 규모와 개방보다 수요를 충족할 수 있는 주체의 데이터의 발굴이 필요하다.

DPV 수치를 종합적으로 보면, 모든 지방자치단체의 DPV 중앙값이 0.4미만이고, 평균값이 중앙값보다 큰 비대칭 분포를 갖고 있다. 이를 DPV 관점에서 해석할 때, 공공데이터포털이 특정한 주제를 검색하고 탐색하는 기능을 제공하고 있는 한편, 사용자는 검색결과에서 선택한 데이터셋의 상세 정보, 데이터셋의 미리보기 기능을 통해 데이터를 실질적으로 확인하는 과정에서 데이터셋을 다운로드하지 않는다고 해석할 수 있다. 이런 현상은 모든 지방자치단체에 공통적으로 나타나기 때문에, 공공데이터포털의 기능 개선이 필요한 부분이다.

4-2 데이터 갱신주기

1) 지자체별 평균 업데이트 경과일

지자체의 데이터 업데이트 경과일은 그림 2에 요약되어 있다. 업데이트 경과일은 서울특별시(1104.4일), 강원도(524.0일), 광주광역시(436.0일), 경상북도(432.3일), 전라남도(419.3일), 세종특별자치시(401.7일), 제주특별자치도(380.9일), 전라북도(354.0일), 경상남도(332.6일), 대전광역시(303.4일), 충청남도(279.3일), 대구광역시(273.5일), 경기도(250.1일), 인천광역시(246.8일), 충청북도(238.4일), 부산광역시(225.5일), 울산광역시(182.8일)의 순서이다. 특히 서울특별시에서 제공한 데이터세트는 가장 낮은 경과일을 갖는 울산광역시에 비해 약 6배 정도 높은 수치이다.

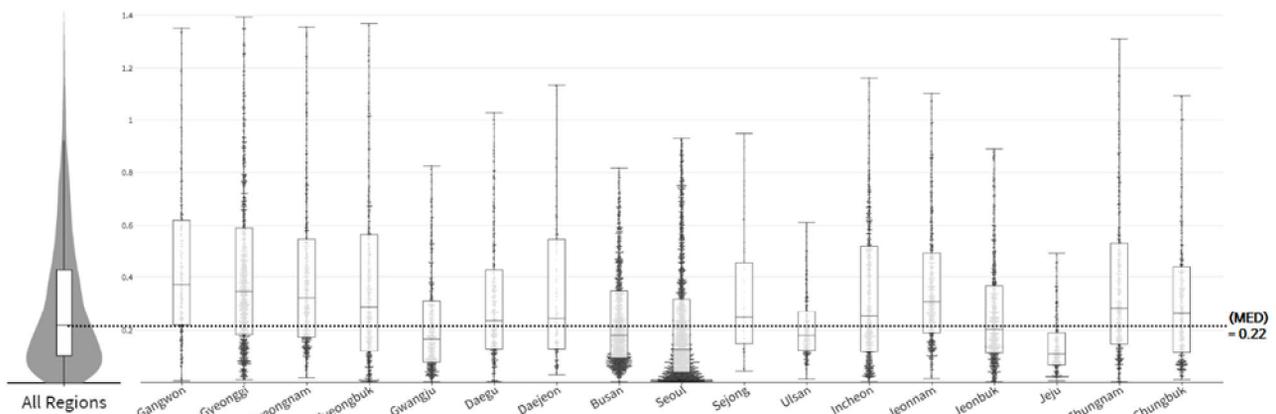


그림 1. 지방자치단체별 DPV 분포
Fig. 1. Distribution of DPV by local governments

표 2. 업데이트 주기별 지방자치단체의 데이터 규모

Table 2. Datasets of local governments by update types

	Annual	Semiannual	Quarter	Monthly	Weekly	Daily	Occasional	null
Seoul	1,524	30	42	39	1	27	968	-
Gyeonggi	1,232	37	30	20	1	3	120	1
Incheon	881	17	9	12		2	36	-
Chungbuk	447	25	10	6			93	2
Chungnam	551	20	6	5	2	5	64	-
Sejong	57	5	2	2				-
Daejeon	251	22	25	10		1	15	13
Jeonbuk	949	23	1	2	1	1	17	-
Jeonnam	642	28	4		1	1	78	3
Gwangju	667	18	7	4			27	1
Jeju	331						29	1
Gyeongbuk	477	11	5	2			210	-
Gyeonnam	664	32	4	9	6	3	49	1
Busan	1,347	170	46	27	3	4	54	1
Ulsan	315	13	3	2	2		6	-
Daegu	493	18	1	10	2	2	36	-
Gangwon	249	10	9	6		5	87	-
N	11,077	479	204	156	19	54	1,889	23

표 3. 카테고리별 업데이트 주기 준수 여부

Table 3. Update Status by categories

Updates	Semiannual	Quarterly	Yearly	Monthly	Weekly	Daily
Timely updates	318	103	8,735	49	-	-
Update pending	161	101	2,342	107	19	54

2) 데이터 갱신 준수율

데이터 갱신 미준수율은 세종특별자치시(48%), 서울특별시(44%), 강원도(44%), 경상북도(38%), 광주광역시(33%), 전라북도(32%), 전라남도(31%), 경상남도(25%), 대구광역시(22%), 경기도(20%), 충청남도(16%), 충청북도(16%), 대전광역시(15%), 부산광역시(11%), 인천광역시(9%), 제주특별자치도(8%), 울산광역시(5%)의 순서로 나타났다. 차기등록예정일의 미준수율은 지자체별로 상이하며, 5%부터 48%까지 특정 수치에 밀집되지 않은 분포를 보인다. 즉 모든 지자체가 차기등록예정일이 준수되지 않은 데이터를 보유하고 있다.

3) 업데이트 주기 준수여부

업데이트 주기의 값은 수집 데이터의 76%에서 '연간' 항목으로 표시되었고, '수시', '반기', '분기'의 순으로 나타났고, 23건은 업데이트 주기가 공백이었다. 표 3은 지방자치단체의 업데이트 주기를 요약하고 있다. 제주특별자치도는 대부분의 데이터를 '연간'으로 업데이트하고 있다.

표 3에서 보듯이, '주간', '일간'으로 표시된 모든 데이터세트는 업데이트 주기를 지키지 못하고 있다. '연간'으로 표시된 데이터세트는 2,342건(21.1%)이 갱신주기를 지키지 못하고 있다. 반면 '분기'로 표시된 데이터세트의 101건(50.5%)이 갱신주기를 지키고 있다. 요약하면, 지방자치단체에서 제공한 데이터세트의 약 20%는 갱신 주기를 준수하지 않고 있다.

그림 3에서 보듯이, 지방자치단체별로 보면 서울특별시가 제공한 데이터세트의 약 42%만 업데이트가 정확한 시점에 진행되었고, 세종특별자치시, 강원도, 광주광역시의 준수율이 각각 53%, 65%, 67% 수준이다. 반면, 제주특별자치도, 울산광역시, 인천광역시, 부산광역시는 90% 이상의 준수율을 갖고 있다. 데이터세트의 수량을 기준으로 보면, 서울특별시, 전라북도, 경기도가 각각 965개, 307개, 222개로 높은 수치를 갖고 있다.

V. 토론과 한계점

공공데이터는 사회 현안을 데이터 기반으로 해결하는 기초가 될 수 있다. 본 연구에서 제안한 데이터 활용지수에 의하면, 지방자치단체에서 개방한 데이터의 활용은 미흡하다. 사용자 관점으로 보면, 검색의 기능적 측면과 데이터세트의 품질 측면으로 개선이 필요하다. 첫째, 사용자의 의도에 맞는 검색 결과가 노출되지 않는다. 공공데이터포털의 검색 기능은 데이터세트의 메타데이터를 활용하고 있는데, 개별 데이터의 특성을 상세하게 기술할 필요가 있다. 일반적인 키워드가 포함된 검색 결과는 재현율(recall)이 높기 때문에, 사용자가 검색 결과를 확인하고 판단하는데 많은 노력이 필요하다. 즉, 조회수가 높지만 데이터 활용지수가 낮은 데이터세트는 사용자의 검색의도와 검색 결과가 일치하지 않는 사례로 해석할 수 있다.

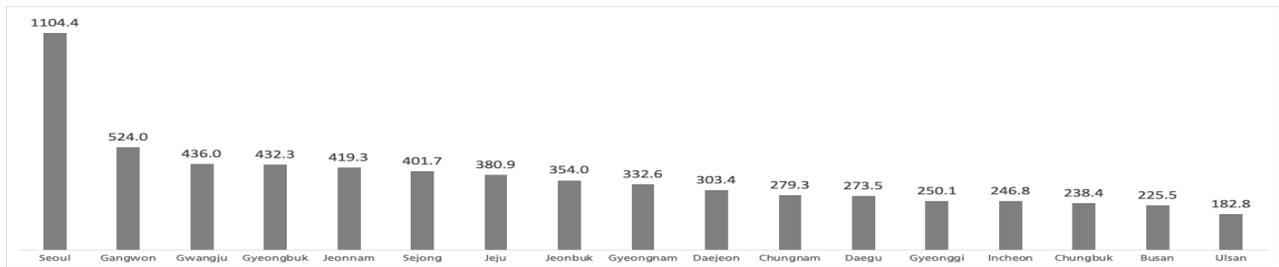


그림 2. 지방자치단체별 평균 업데이트 경과일
 Fig. 2. Elapsed update days by local governments (average)

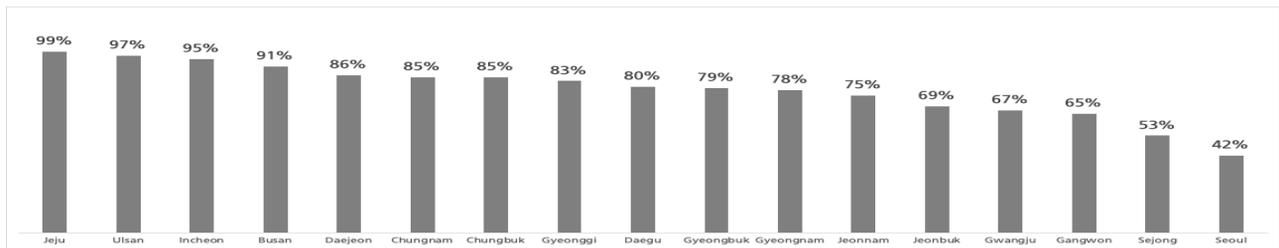


그림 3. 업데이트 준수 여부
 Fig. 3. Average rate about updates of datasets on time

이런 문제의 해결을 위해, 데이터 관점에서 데이터세트의 메타데이터를 정확하고 정교하게 표현해야 하고, 메타데이터의 의미 정보를 활용한 검색 기능을 시스템 측면에서 제공해야 한다. 둘째, 개별 데이터세트가 사용자 의도에 맞는 정보를 포함하지 않을 수 있다. 공공데이터포털은 데이터세트에 대해 미리 보기 기능을 제공하고 있다. 사용자는 특정한 데이터세트를 선택하고, 데이터세트가 포함한 메타데이터 (컬럼)와 데이터의 사례를 확인할 수 있다. 즉, 사용자는 데이터를 확인하고, 활용 여부를 결정할 수 있기 때문에, 조회 이후 다운로드되지 않은 데이터세트는 사용자의 의도와 다른 데이터로 해석할 수 있다. 수요 중심의 데이터 발굴과 제공은 이 문제를 해결하는 주요한 방안으로 검토할 수 있다. 사용자가 원하는 데이터를 조사하고, 적시에 제공할 수 있는 체계를 검토할 필요가 있다. 특히, 지방자치단체는 데이터 개방에 필요한 인력과 시스템이 부족할 수 있기 때문에, 공공데이터의 개방과 활용 체계를 개선할 필요가 있다[24].

데이터 사용자 관점으로 보면, 데이터포털은 데이터의 접근성, 데이터 어휘 표준, 데이터 관리 방안을 개선해야 한다. 먼저, 일관성 있는 데이터의 접근성을 확보해야 한다. 다수의 데이터 세트는 공공데이터포털이 아닌 데이터 제공기관의 연결정보를 제공한다. 공공데이터포털이 ‘포털’의 기능을 제공하기 위해 데이터의 제공 방식을 검토해야 한다. 특히, 연결된 사이트나 서비스에 따라 해당 데이터에 접근하지 못하는 사례도 있기 때문에 이에 대한 개선이 필요하다. 둘째, 공공데이터포털과 지방자치단체의 분류체계, 메타데이터 항목에 대한 표준이 필요하다. 예를 들어, 공공데이터포털은 주제 분류를 ‘공공행정’, ‘과학기술’, ‘교육’, ‘교통물류’, ‘국토관리’, ‘농축수산’, ‘문화관광’, ‘법률’, ‘보건의료’, ‘사회복지’, ‘산업고용’, ‘식품건강’, ‘재

난안전’, ‘재정금융’, ‘통일외교안보’, ‘환경기상’ 등 16개로 구분하고 있다. 반면, 서울특별시의 열린데이터광장은 ‘보건’, ‘일반행정’, ‘문화/관광’, ‘산업/경제’, ‘복지’, ‘환경’, ‘교통’, ‘도시관리’, ‘교육’, ‘안전’, ‘인구/가구’, ‘주택/건설’ 등 12개로 구분한다. 그러나 주제 분류명이 다르고, 포함하는 주제가 서로 다르기 때문에 데이터를 일관성 있게 해석하는 것이 어렵다. 셋째, 공공데이터포털, 지방자치단체의 데이터포털에 공개한 모든 데이터를 덤프 (dump) 파일로 접근할 수 있도록 제공하는 것을 검토해야 한다. 현재의 데이터포털은 사용자가 소수의 데이터를 특정해서 다운로드받는 것을 핵심 기능으로 갖고 있다. 그러나, 다수의 데이터세트나 데이터포털의 대규모 데이터를 활용하기 위해 개별적인 검색과 다운로드는 효과적이지 않다. 따라서, 주기적으로 전체 데이터세트를 압축하고, 접근할 수 있는 인터페이스를 제공하는 방안을 검토해 볼 필요가 있다. 이런 방식은 운영과 관리 측면에서 이슈가 있지만, 공공데이터의 활용 활성화를 위한 방안이다.

본 연구는 공공데이터포털에서 제공하는 데이터의 활용 여부를 정량적으로 측정하고 있다. 그러나, 데이터 수집과 분석에 있어 다음의 한계점을 갖고 있다. 첫째, 공공데이터포털과 지방자치단체에서 운영하는 데이터 포털의 관계를 고려하지 않았다. 본 연구에서 수집한 데이터세트는 공공데이터포털에서 제공하는 것에 한정되기 때문에, 자체적으로 데이터포털을 운영하는 지방자치단체의 데이터세트와 일치하지 않을 수 있고, 데이터 활용지수도 다를 수 있다. 둘째, 데이터 활용지수 (DPV)는 조회수와 다운로드의 관계를 정의한 것으로 실질적인 데이터 활용 여부를 평가하지 않는다. 즉, 사용자가 공공데이터포털의 데이터를 탐색과 접근한 정도를 측정할 수 있으나, 다운로드된 데이터의 활용 여부는 판단하기 어렵다. 그럼에도, 사용자가 다

온로드한다는 것은 데이터 활용을 위한 전제조건이기 때문에 데이터 활용지수를 통해 간접적으로 해석이 가능하다. 셋째, 갱신 주기의 준수 여부는 등록일과 업데이트 주기를 고려해 분석되었다. 그러나 데이터세트의 정확한 등록일자를 확인할 수 없기 때문에, 분석 결과는 실제 갱신주기보다 높게 평가될 수 있다.

VI. 결 론

본 연구는 지방자치단체에서 공공데이터포털에 개방한 데이터의 활용성을 평가하고 있다. 지방자치단체에서 개방한 공공데이터의 주제 분류와 규모는 서로 다르지만, 데이터 활용은 전반적으로 미흡한 수준이다. 데이터 활용지수를 기준으로 보면, 지방자치단체가 개방한 데이터 규모에 관계없이 활용성은 미흡한 수준이다. 지방자치단체에 따라 개방한 공공데이터의 규모는 다르지만, 활용지수는 큰 차이 없이 낮게 나타났다. 공공데이터포털에서 특정한 주제의 데이터세트를 검색하고 확인하는 결과는 데이터세트의 조회와 다운로드 수치로 확인할 수 있다. 그러나, 데이터세트의 조회보다 다운로드가 비교적 낮기 때문에, 데이터의 검색이 사용자의 활용으로 나타나지 않았다고 해석할 수 있다. 이러한 결과는 전체 지방자치단체에서 공통적으로 나타났다. 데이터의 업데이트 주기는 관리 측면에서 평가할 수 있는 항목이다. 분석에 따르면 전체 데이터세트의 약 20%가 업데이트 주기를 준수하지 않고 있었다.

공공데이터의 활용 활성화는 법제도, 관리체계, 데이터 품질 등 다양한 관점의 균형 있는 접근이 필요하다. 특히, 검색 기능은 사용자가 의도에 맞는 데이터세트를 탐색하고, 활용 여부를 판단하는데 중요한 수단이다. 데이터세트의 메타데이터는 연관 있는 정보를 정확하고 풍부하게 기술해야 하고, 분류체계를 일관성 있게 적용할 필요가 있다. 한편, 데이터의 개방 확대와 수요를 반영하는 데이터의 발굴이 균형 있게 진행되어야 한다.

본 연구는 지방자치단체에서 제공한 공공데이터 활용도를 공공데이터포털에 한정해서 측정했다. 지방자치단체가 자체적으로 데이터포털을 운영하는 사례가 있기 때문에, 향후 연구는 지방자치단체의 데이터 포털에 개방된 데이터를 대상으로 활용도를 평가하는 것이 필요하다. 더불어, 지방자치단체의 데이터 포털과 공공데이터포털의 데이터 연계와 시스템 사이의 연동 여부를 검토함으로써 공공데이터의 전반적 활용도를 평가하는 방안에 대한 연구가 필요하다.

참고문헌

[1] Ministry of the Interior and Safety. ACT ON PROMOTION OF THE PROVISION AND USE OF PUBLIC DATA. Available: <https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2§ion=lawNm&query=공공데이터&x=0&y=0#liBgcolor2>

[2] Ministry of the Interior and Safety. Guidelines of Public Data. Available: https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000016&nttId=72733

[3] H. L. Kim, "Metadata Analysis of Open Government Data by Formal Concept Analysis," Journal of the Korea Contents Association, Vol. 18, No. 1, pp. 305-313, January 2018

[4] H. J. Seo, and S. H. Myeong, "Policy Alternatives for Public Data Disclosure by Local Government : Focusing on Operation of Public Data Portal Site," Journal of Korean Association for Regional Information Society, Vol. 18, No. 4, pp. 1-27, December 2015

[5] S. G. Kim, and J. H. Cho, "Proposal for the introduction of big data by local governments," Journal of Korean Association for Regional Information Society, Vol.16, No.3. September 2013

[6] V. Weerakkody, Z. Irani, K. Kapoor, U. Sivarajah, and Y. K. Dwivedi, "Open data and its usability: an empirical view from the Citizen's perspective," Information Systems Frontiers, Vol. 19, pp. 285-300, April 2017

[7] H. J. Seo, "Discussion for Securing Self-Reliant Finance of Local Government Using Public Data," Journal of the Korea Contents Association, Vol. 15, No. 4, pp. 380-390, April 2015

[8] J. H. Kim, S. W. Cho, K. H. Lee and W. S. Cho, "Service Level Evaluation Through Measurement Indicators for Public Open Data," Korea Bigdata Society, Vol. 1 No. 1, pp. 53-60, 2016

[9] J. E. Kang, "A legal study on open data," Administrative Law Journal, Vol. 61, pp. 159-178, May 2020

[10] Y. W. Hong, "A study on the invigorating strategies for open government data," Journal of the Korean Data And Information Science Society, Vol. 25, No. 4, pp. 769-777, August 2014

[11] F. KIRSTEIN, B. DITTWALD, S. DUTKOWSKI, Y. GLIKMAN, S. SCHIMMLER, and M. HAUSWIRTH, "Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP," in International Conference on Electronic Government, EGOV 2019, pp. 192-204, 2019

[12] R. Machova et al., M. Hub, and M. Lnenicka, "Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective," Aslib J. Inf. Manag., Vol. 70, No. 3, pp. 252-268, May 2018

[13] S. H. Kim, C. S. Lee, S. H. Chung, H. C. Kim and C. S. Lee, "A Study of Public Data Quality Management Maturity Model," Proceedings of the Korean Institute of Industrial Engineers Fall Conference, Suwon, pp. 2220-2230, November 2014

- [14] G. E. Park and C. J. Kim, "Quality Characteristics of Public Open Data," The Korea Society of Digital Policy & Management, Vol. 13, No. 10, pp. 135-146, 2015
- [15] H. L. Kim, "Quality Evaluation of the Open Standard Data," Journal of the Korea Contents Association, Vol. 20, No. 9, pp. 439-447, September 2020
- [16] P. Archer, Data Catalog Vocabulary (DCAT), W3C Recommendation, Online, 2014, Available: <https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/>
- [17] S. Neumaier, J. Umbrich, and A. Polleres, "Automated Quality Assessment of Metadata across Open Data Portals," ACM J. Data Inf. Qual., Vol. 8, No. 2, pp.1-29, October 2016
- [18] P. Kremen, and M. Necaský, "Improving discoverability of open government data with rich metadata descriptions using semantic government vocabular," J. Web Semant., Vol. 55, pp. 1-20, March 2019
- [19] J. Klímek, "DCAT-AP representation of Czech National Open Data Catalog and its impact," J. Web Semant., Vol. 55, pp. 69-85, March 2019
- [20] J. Klímek, and P. Skoda, "LinkedPipes DCAT-AP Viewer: A Native DCAT-AP Data Catalog," in International Semantic Web Conference, 2018
- [21] E. J. Kim, M. S. Kim, and H. W. Kim, "Data Standardization for the Enhanced Utilization of Public Government Data," Knowledge Management Research, Vol. 20, No. 4, pp. 23-38, 2019.
- [22] J. W. Gim, D. W. Jeong, J. H. Lee, S. J. Lim and J. Y. Kim, "A Study on Data Quality Evaluation Factors Considering Big Data Environment," in Proceedings of KIIT Conference, pp. 258-259, December 2017
- [23] S. H. Song, J. Y. Kim and J. Y. Lee, "A Study on Public Open Data Status and Activation Plan for Use," in Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 854-855, June 2013
- [24] H. J. Song, and S. S. Hwang, "Seeking Strategies for Local Governments to Prepare for Public Data Act," Journal of Korean Association for Regional Information Society, Vol. 17, No. 2, pp. 1-28, June 2014

김동준(Dongjun Kim)



2018년 : 중앙대학교 문헌정보학과 (재학)

2018년~현재 : 중앙대학교 문헌정보학과 (재학)

※관심분야 : 정보격차, 메타데이터, 소셜 인포매틱스 등

김현지(Hyeonji Kim)



2021년 : 중앙대학교 문헌정보학사

2016년~2021: 중앙대학교 문헌정보학과

※관심분야 : 전자정보, 공공데이터, 디지털 아카이브 등

송채은(Chaeun Song)



2016년 : 중앙대학교 문헌정보학과 (재학)

2016년~2021: 중앙대학교 문헌정보학과

※관심분야 : 데이터 사이언스, 공공데이터, 지식그래프 등

양지우(Jisoo Yang)



2020년 : 중앙대학교 문헌정보학사

2016년~2020년: 중앙대학교 문헌정보학과

※관심분야 : 지식그래프, 공공데이터, 소셜네트워크 등

김학래(Haklae Kim)



2010년 : 아일랜드 국립대학교 (공학박사)

2004년~2009년: Digital Enterprise Research Institute, Ireland

2009년~2016년: 삼성전자

2017년~2019년: 한국과학기술정보연구원

2019년~현재 : 중앙대학교 문헌정보학과 교수

※관심분야 : 지식그래프, 인공지능, 데이터 사이언스 등