한국빅데이터학회지
제1권 제1호(창간호), 2016, pp. 53-60

공공데이터 개방 평가지표 개발을 통한 현황분석 및 가시화*

Service Level Evaluation Through Measurement Indicators for Public Open Data

김지혜[1†] · 조상우[2] · 이경희[2] · 조완섭[1]

충북대학교 대학원 경영정보학과[1]
충북대학교 비즈니스데이터융합학과[2]

요 약

공공데이터 포털에 공개된 지자체 데이터와 공공기관 데이터를 자동으로 수집한 후, 공공 데이터의 개방현황 및 다양한 영역별 데이터 제공여부, 파일 형식 등 다양한 기준으로 다차원 분석하여 서비스 수준 평가를 제공하고자 한다. 이를 위해, 해외 평가지표 사례를 바탕으로 평가지표 내용을 설정한 후 이를 기준으로 데이터웨어하우스(DW)를 구축하였으며, 다차원 분석 기법을 사용한 서비스 수준평가 결과를 지역별로, 기관별로, 분야별로 시각화한다.

■ 중심어 : 빅데이터, 공공데이터, 평가지표, 다차원 분석, 자연어 처리방식, 시각화

Abstract

Data of central government and local government was collected automatically from the public data portal. And we did the multidimensional analysis based on various perspective like file format and present condition of public data. To complete this work, we constructed Data Warehouse based on the other countries' evaluation index case. Finally, the result from service level evaluation by using multidimensional analysis was used to display each area, establishment, fields.

■ Keyword : Big Data, Open Data, Evaluation, Multidimensional Analysis, NLP, Visualization

Ⅰ. 서 론

최근 공공데이터 포털(data.go.kr)을 통해 공개한 정부기관 및 지자체의 공공데이터를 활용한 서비스가 증가하고 있는 추세이다. 그와 더불어 정부 및 지방자치단체는 공공데이터 제공을 위한 정책 수립과 운영에는 익숙하지만, 현재까지 공공데이터개방 평가지표는 제시하지 않은 상태이다.

본 논문에서는 '개방된 데이터의 다양한 민간 활용'을 위해서 공공데이터 포털에 공개된 데이터를 기반으로 하여 각 지역별, 기관별 공공데이터의 현황을 좀 더 세부적으로 평가하기 위해 공공데이터개방 평가지표를 개발하고, 개발한 평가지표를 기준으로 지자체 및 공공기관의 공공데이터 개방지수를 산정한다. 이와 같은 공공데이터 평가지표는 공공데이터 개방시 데이터의 품질, 활용에 편리한 데이터 포맷 제공 및 데이터의 지속적인 업데이트를 유도함으로써 공공데이터를 활용한 고품질 공공서비스 개발에 초석을 제공할 수 있을 것으로 기대한다.

공공데이터 개방 평가지표를 산정하는 시스템을 구축하기 위하여 공공데이터 포털(data.go.kr)에 개방된 데이터를 수집하여 DW를 구축한다. 구축한 DW는 각 공

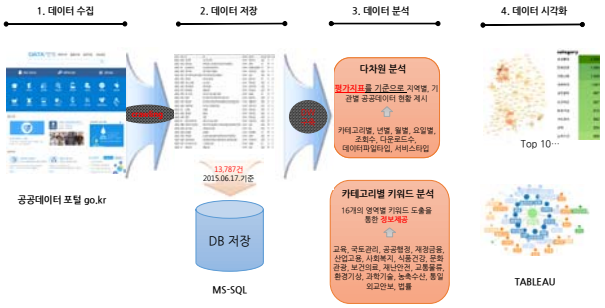
2015년 12월 12일 접수; 2016년 1월 8일 수정본 접수; 2016년 1월 15일 게재 확정.
* 본 논문은 미래창조과학부의 2015년 고용계약형SW석사과정지원사업의 지원을 받았음(H0116-15-1011)
† 교신저자 kjh-wisdom@hanmail.net

공기관별, 지방자치단체별로 공개한 데이터의 건수, 파일형식, 데이터 주제영역, 조회수와 다운로드 수 등을 명확하게 파악할 수 있으며, 이를 토대로 공공데이터 평가지표를 산정할 수 있다.

Ⅱ. 본 론

2.1 전체 연구 프로세스

설계된 본 논문의 연구 프로세스는 <그림 1>과 같다. 본 연구는 2012년 12월 1일부터 2015년 6월 17일까지의 데이터를 사용한다. 수집된 데이터는 기관별로 식별이 가능하다. Java를 통해 데이터 정제를 한 뒤 DB를 구축하였다. 그리고 다차원분석을 위해 DW를 구축하여 분석한 후 그 결과를 시각화 하였다.



〈그림 1〉 연구전체 프로세스

본 연구에서는 공공데이터 포털에서 데이터를 수집하는 웹크롤러를 개발하여 주기적으로 데이터를 수집한다. 수집한 데이터 구조는 <표 1>과 같다.

〈표 1〉 수집한 데이터 구조

연번	컬럼명		설명
	영어	한글	
1	data_type	테이터타입	파일데이터, 오픈API, Visual
2	category	영역별	16개의 영역별
3	title	제목	공개 테이터 이름
4	body	내용	공개데이터 세부 내용
5	modify	수정일	공개데이터 올린 날짜
6	organization	기관	8개 기관별
7	service_type	제공 서비스타입	
8	file_type	파일형식	파일 형태 표시
9	view_count	조회수	사용자가 열어본 횟수
10	download	다운로드수	사용자가 직접 다운받은 횟수

데이터 전처리 과정에서 기관을 기관 1, 기관 2로 더 세분화한 후 데이터 데이터베이스에 저장한다. 그 다음 다차원분석을 위해 DW를 구축하고, OLAP 분석한 결과를 QGIS와 Tableau를 통해 시각화 한다.

2.2 평가지표 설정

2.2.1 Global Open Data Index

Global Open Data Index는 글로벌 데이터인덱스포럼에서 만든 공공데이터 평가지표로서, 매년 정부가 개방형태(open format)로 만들고 있는 데이터들이 실제로 시민, 언론, 시민사회단체 등에서 접근가능하고 잘 활용되고 있는지를 개방형 데이터 전문가들에게 설문한 결과를 기반으로 평가지수를 산정하고, 공개한다 [http://global.census.okfn.org/]. 글로벌데이터인덱스를 참고하여 공공데이터 개방 평가지표를 도출하였다. 개발한 평가지표는 각각의 데이터 셋은 9개의 서로 다른 질문을 통해 평가하며, 기술적 측면과 법적 측면의 두 가지 측면을 중요한 축으로 하였다. 이를 바탕으로 4개의 세부내용과 분석방법은 <표 2>와 같다.

〈표 2〉 평가지표 내용 및 분석방법

지표	세부내용	분석방법
1. 국민에게 제공할 의무에 따른 데이터 개방 현황	가장 기본이 되어야 할 법적 의무에 따른 시행 여부를 평가함.	각 지역별, 기관별 데이터 현황 시각화: 지역별(시도, 시군구)/기관별
2. 개방된 파일형식에 있어 가독성 현황	데이터가 기계가독형 형태인지를 평가하기 위한 것으로 데이터는 디지털형태이지만 PDF문서처럼 기계가독형이 아닌 것들이 존재함.	Five Star Open Data 기준에 따른 현황 시각화 1 유형 pdf, jpg, png 2 유형 TXT-TEXT, hwp, XLS-XLSX-EXCEL, 3 유형 csv, xml, 4 유형 uri 5 유형 rdf, docx: 지역별(시도, 시군구)/기관별
3. 개방데이터의 다운로드 수에 따른 실제 활용 현황	실제 다운로드 수와 조회수 비교를 통한 데이터 질과 관련된 실제 사용에 미친 영향력을 평가함.	각 지역별, 기관별 데이터 다운로드 수와 조회수 현황 비교를 통해 실제 사용되어지는 데이터를 파악: 지역별(시도, 시군구)/기관별
4. 다양한 영역의 데이터 제공 현황	데이터의 다양성을 평가하는 것으로 각 지역별로 다양한 영역의 데이터를 제공하고 있는지를 평가함.	다차원분석을 통해 파악 후 시각화: 지역별(시도, 시군구)

그 첫 번째로 국민에게 제공할 공공데이터 개방의 의무성과 개방된 파일 형식 상태 파악으로 테이터 타입별 평가의 필요성을 인식하고 공공데이터 개방 평가지표에 참고하였다.

2.2.2 Five Star Open Data

웹, 데이터의 창시자인 팀 버너스리가 별점을 활용하여 개방형 데이터의 단계(수준)를 설명하고 각 단계별 비용과 효과를 설명하였다. 각 단계별 효과 및 특징은 다음 <표 3>과 같다.

〈표 3〉 5 Star Open Data의 내용

유형	형식	종류	특징
1	OL	표를 스캔한 이미지 파일 pdf, jpg, png	인쇄가능, 하드 드라이브나 USB 저장 가능, 타 시스템에 수작업 입력 가능
2	OL, RE	표를 엑셀파일과 같은 특정 포맷으로 공개 TXT-TEXT, hwp, XLS-XLSX-EXCEL	데이터 수집, 시각화 등을 특정 소프트웨어를 사용해 처리가능, 다른 포맷으로 데이터 발행 가능
3	OL, RE, OF	표를 엑셀 대신 CSV파일과 같은 비독점 포맷으로 제공 csv,xml	특정 소프트웨어 기능에 한정되지 않고, 원하는 방법으로 데이터 작업가능
4	OL, RE, OF, URI	표 값을 구분하기 위해 URI 부여	로컬 및 웹상에 링크가능, 북마크 가능, 데이터의 일부를 재사용 가능
5	OL, RE, OF, URI, LD	다른 데이터 소스와의 링크를 통해 연관지식 획득 및 활용효과 극대화 가능 RDF, docx	관련 데이터를 찾을 수 있으며 데이터 스키마에 대해 직접 배울 수 있음

* OL = 오픈 라이센스(Open Licence), RE = 기계가독형(Machine Readable), OF = 개방형 포맷(Open Fomat), URI = 개체식별을 위해 URI를 사용, LD = 링크드 데이터(Linked Data)

이 지표의 특징은 Open Licence를 제외한 기계가독형, 개방형 포맷, URI, Linked Data는 평가지표로 실제 개방된 데이터를 직접적으로 확인할 수 있는 기술적인 활용 가능하다.

2.3 지역별 기관별 분류

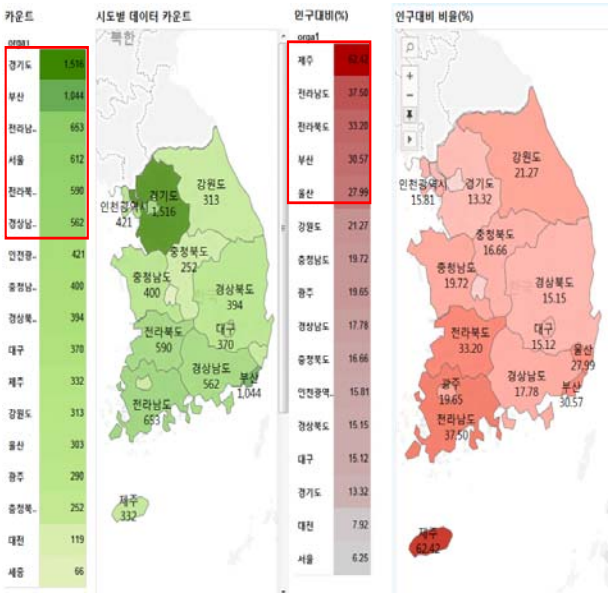
본 연구 분석에 있어 지역별(자치행정조직), 기관별로 <표 4>와 같이 나누어 평가하였다.

〈표 4〉 지역별과 기관별 분류

지역별		기관별(7)	기관수(354)
시도수(17)	시군구(232)		
강원도	15	공공기관	285
경기도	30	교육기관	4
경상남도	18	교육행정조직	17
경상북도	24	국가행정기관	43
전라남도	23	위원회 및	2
전라북도	14	경제자유구청	1
충청남도	15	입법조직	2
충청북도	11	헌법조직	
광주광역시	7		
대구광역시	10		
대전광역시	6		
부산광역시	16		
서울특별시	22		
세종특별시	0		
울산광역시	5		
인천광역시	13		
제주특별도	2		

2.4 지역별 분석결과

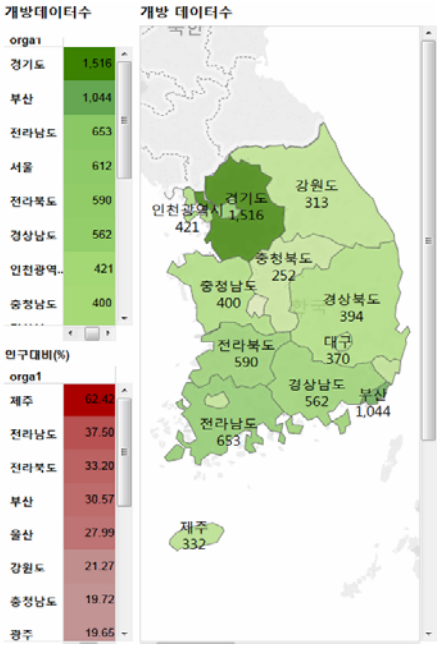
지역의 총 데이터 수는 8,237건이었다. 그 중에 각 지역별 공공데이터 개방 현황은 <그림 2>와 같다.



〈그림 2〉 전국 공공데이터 개방 현황 결과

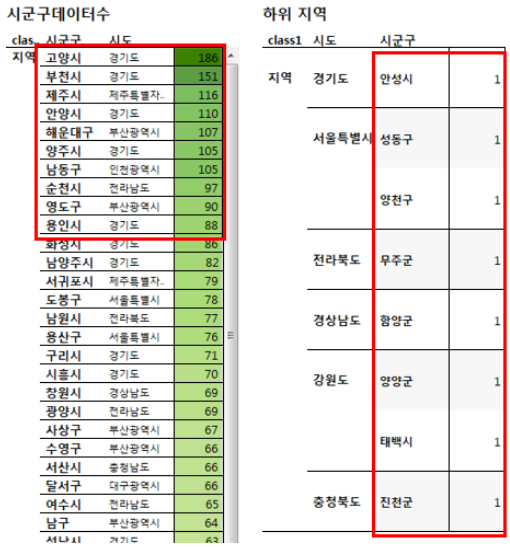
2.4.1 국민에게 제공할 의무에 따른 공공데이터 개방현황

평가지표 1번 결과에 시도별 공공데이터 개방현황은 <그림 3>와 같다. 인구대비 공공데이터 개방 수를 비교하였을 때 가장 많은 데이터를 올린 지역은 제주도였으며 전라남도, 전라북도, 부산 순으로 볼 수 있었다.



〈그림 3〉 공공데이터 개방 평가지표 1번 결과_시도

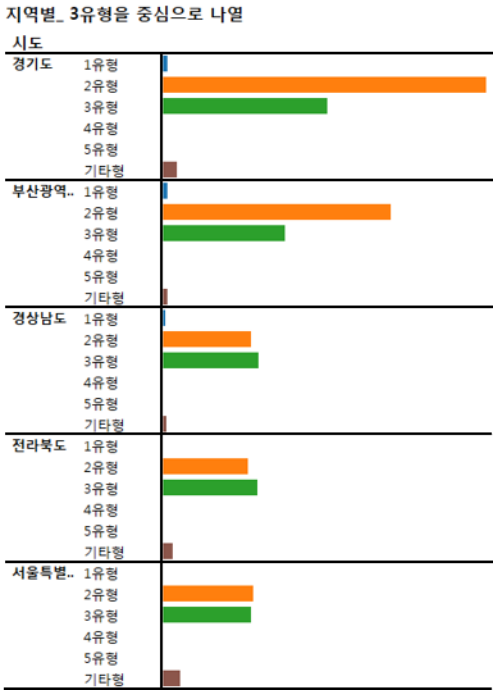
시군구별 공공데이터 개방 현황은 <그림 4>와 같다. 시군구별로 공공데이터 개방 현황을 파악했을 때 제주시가 인구대비 가장 많은 데이터 수를 제공해 주었으며 해운구, 순천시, 영도구 순으로 나타났다.



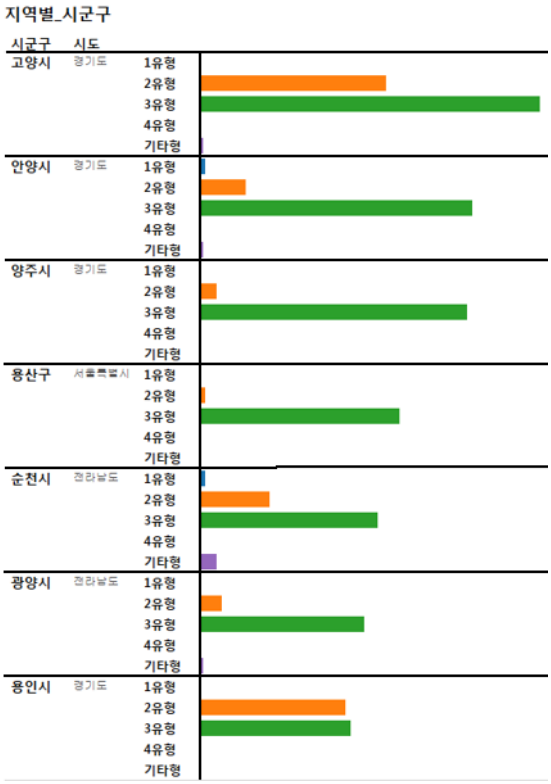
〈그림 4〉 공공데이터 개방 평가지표 1번 결과_시군구

2.4.2 개방된 데이터의 파일 형식에 따른 데이터 가독성 현황

평가지표 2번 결과에 따른 시각화는 <그림 5>, <그림 6>과 같다.



〈그림 5〉 공공데이터 개방 평가지표 2번 결과_시도



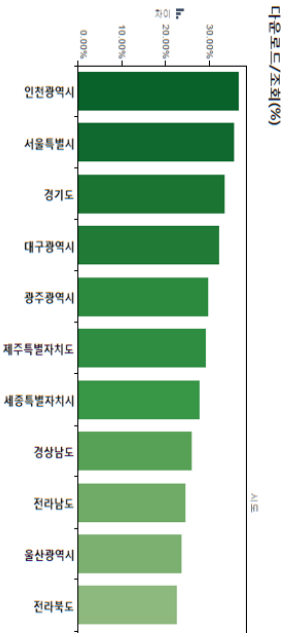
〈그림 6〉 공공데이터 개방 평가지표 2번 결과_시군구

평가지표 1번에서는 각 데이터 개방 자체에 대한 현황에 대한 단순한 카운트에 대한 인구대비 비율이었다면, 평가지표 2에서는 좀 더 세부적으로 들어가 공개데이터 품질에 있어 지역전체의 공공데이터 형식은 Five Star Open Data 기준인 4유형(uri), 5형식(rdf, docx)의 데이터 형식은 0에 가까웠다. 3유형을 우선순위로 공공데이터 개방 현황을 봤을 때 가장 많이 공개된 지역(시군구)은 고양시로 나타났으며 다음으로는 안양시, 양주시, 용산구 순으로 나타났다.

2.4.3 개방된 공공데이터 활용 여부 현황

데이터의 다운로드 수와 조회수의 비교를 통해 실제 조회수는 많지만 다운로드 수가 얼마나 많은지 비교해 보았다. 조회수가 많다는 것은 사람들의 관심이 많다는 의미이고 실제 다운로드 수가 많다는 것은 데이터의 질에 있어서 사용하기 쉽다는 것을 추측하여 둘을 비교하여 보았다.

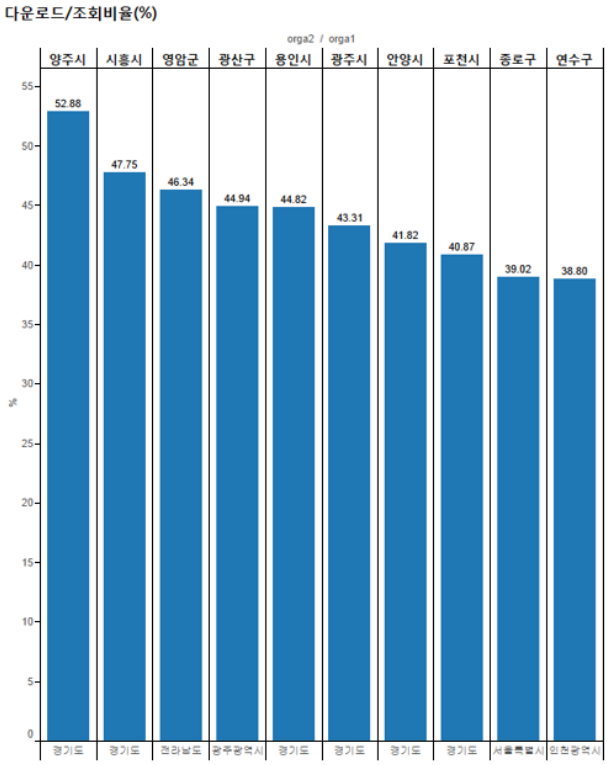
그 결과 평가지표 3에 따른 공공데이터 개방 시도현황은 <그림 7>과 같다.



〈그림 7〉 공공데이터 개방 평가지표 3번 결과_시도

조회수 대비 다운로드수가 가장 많은 지역은 인천광역시로 결과가 나타났다. 경기도 같은 경우 조회수가 최고로 많았지만 다운로드 수는 적은 것을 볼 수 있었다.

다음은 시군구별 공공데이터 개방 현황은 <그림 8>과 같다.



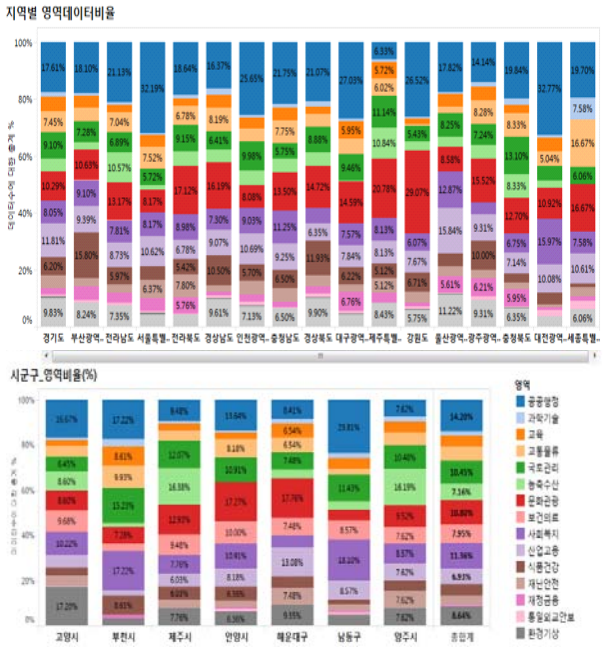
〈그림 8〉 공공데이터 개방 평가지표 3번 결과_시군구

조회수는 부천시, 양주시, 안양시, 고양시 순으로 나타났지만 실제 조회수 대비 다운로드 수의 결과는 양주시 52%, 시흥시 47%, 영암군 46%순으로 나타났다.

2.4.4 지역별 다양한 영역(16개) 데이터 개방 현황

다음은 각 지역별 다양한 영역(공공행정, 문화관광, 산업고용, 사회복지, 교통물류, 보건의료, 환경기상, 국토관리, 교육, 농축수산, 과학기술, 재난안전, 식품건강, 재정금융, 통일외교안보, 법률)과의 다차원 분석을 통해 데이터의 다양성을 평가한 결과이다. 평가지표 4번에 대한 시각화는 <그림 9>와 같다.

경기도 지역이 법률 외에 15개 영역에서 개방된 데이터를 제공하고 있었으며 부산, 전라남도, 서울특별시 순을 볼 수 있었다. 또한 대부분의 데이터가 공공행정 부분에 치우쳐 있었으며 문화 관광 그리고 각 지역별 특징에 따라 산업고용 혹은 농축수산, 사회복지 등으로 데이터 양의 변화를 파악할 수 있었다. 예를들어 서울특별시에 경우 공공행정 분야가 32%로 가장 많이 차지하였으며 강원도의 경우 문화 관광 데이터가 29%로 많은 비율을 차지하고 있었다. 부산광역시의 경우에는 다양한 골고루 분포되었으나 산업고용 영역의 데이터가 다른 지역에 비해 15%로 가장 높았다.



〈그림 9〉 공공데이터 개방 평가지표 4번 결과

따라서 지역별 공공데이터 개방 평가지표 현황 결과, 시도별 Top5는 다음 <표 9>와 같다.

〈표 9〉 공공데이터 개방 평가지표 전체결과_시도

순위	평가지표 1	평가지표 2	평가지표 3	평가지표 4
1	전남	경기	인천	경기
2	전북	부산	서울	부산
3	부산	경남	대구	서울
4	경기	전남	경기	전남
5	서울	전북	광주	전북

공공데이터 개방 평가지표에 따른 종합 순위는 각 평가지표에 따라 만점을 기준으로 평가하에 경기도가 37.5점으로 가장 높았으며 부산, 서울, 경상남도, 전라남도로 Top5를 도출하였다. 다음은 시군구별 Top5는 다음 <표 10>과 같다.

〈표 10〉 공공데이터 개방 평가지표 전체결과_시군구

순위	평가지표 1	평가지표 2	평가지표 3	평가지표 4
1	고양시	고양시	양주시	고양시
2	부천시	안양시	시흥시	부천시
3	제주시	양주시	영암군	제주시
4	안양시	용산구	광산구	안양시
5	해운대구	순천시	용인시	해운대구

종합점수 1위 고양시의 경우는 평가지표 1번 인구대비 공공데이터수 비율에서는 점수를 얻지 못했지만 나머지 평가지표 2번인 데이터의 형식면 그리고 평가지표 3번 조회수 대비 다운로드 건이 많았으며 평가지표 4번인 지역별 다양한 영역에서 개방된 데이터가 골고루 분포되어 있었다.

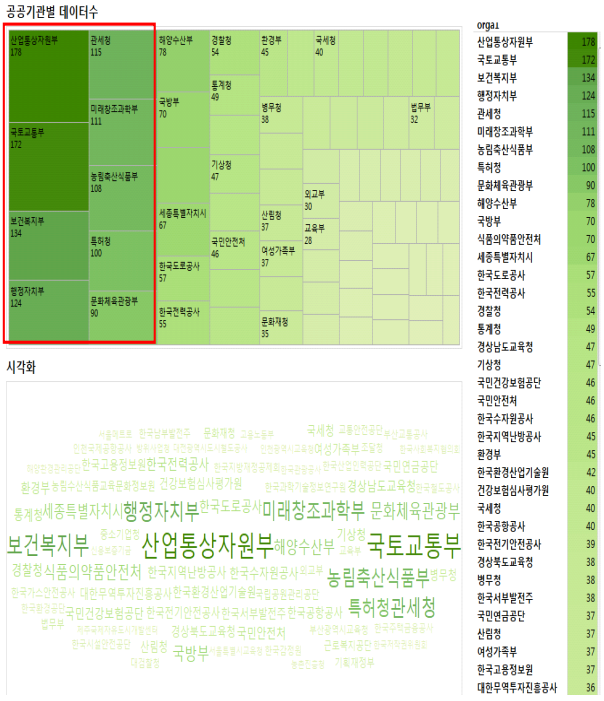
2.5 기관별 분석결과

기관의 총 데이터 수는 5,550건이었다. 이 중에 10개의 기관별 공공데이터의 경우 사법조직과 경제자유구역청(조합), 기타 기관을 제외한 공공기관(3261), 교육기관(8), 교육행정조직(267), 국가행정기관(2455), 위원회 및 경제자유구청(21), 입법조직(11), 헌법조직(13) 7개의 기관으로 분류하여 세부기관별 현황을 파악하였다.

2.5.1 국민에게 제공할 의무에 따른 공공데이터 개방현황

공공데이터 활용지원센터 533건으로 가장 높은 것을 제외한 나머지를 다음과 <그림 10>과 같이 비교하여 현황을 파악하였다.

그 결과 산업통상자원부, 국토교통부, 보건복지부, 행정자치부, 관세청, 미래창조과학부, 농림축산식품부, 특허청, 문화체육관광부 등의 순서로 데이터가 많았다.

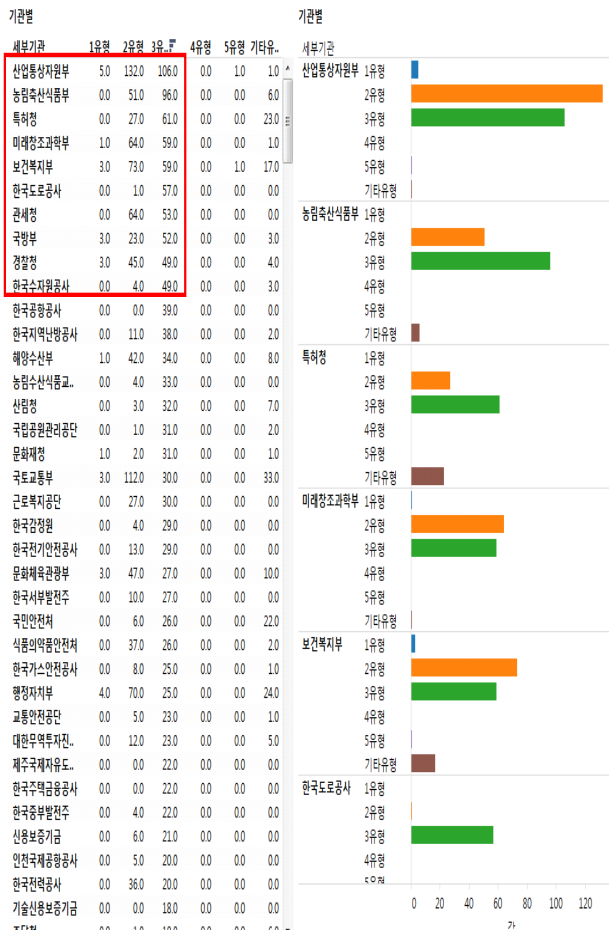


〈그림 10〉 공공데이터 개방 평가지표 1번 결과

2.5.2 개방된 데이터의 파일 형식에 따른 데이터 가독성 현황

결과 <그림 11>과 같았다. 3유형으로 기준으로 데이터 수를 비교했을 때 가장 많이 올린 지역이 산업통상자원부이며 농림축산식품부, 특허청 순이었다. 그러나 대부분 2유형을 올린 기관은 미래창조과학부를 비롯한 보건복지부와 관세청 데이터가 높은 것을 볼 수 있었다.

다음은 데이터 수가 1개로써 1~5유형에도 속하지 않는 기타형 한글문서 HWP등 보고서와 같은 데이터를 올린 기관은 게임물관리위원회, 대한장애인체육회, 세종학당재단, 에너지경제연구원, 육아정책연구소, 한국법제연구원, 한국지방행정연구원, 한국해양대학교로 나타났으며, 그 밖의 데이터들은 보통 2유형을 따르고 있었다.

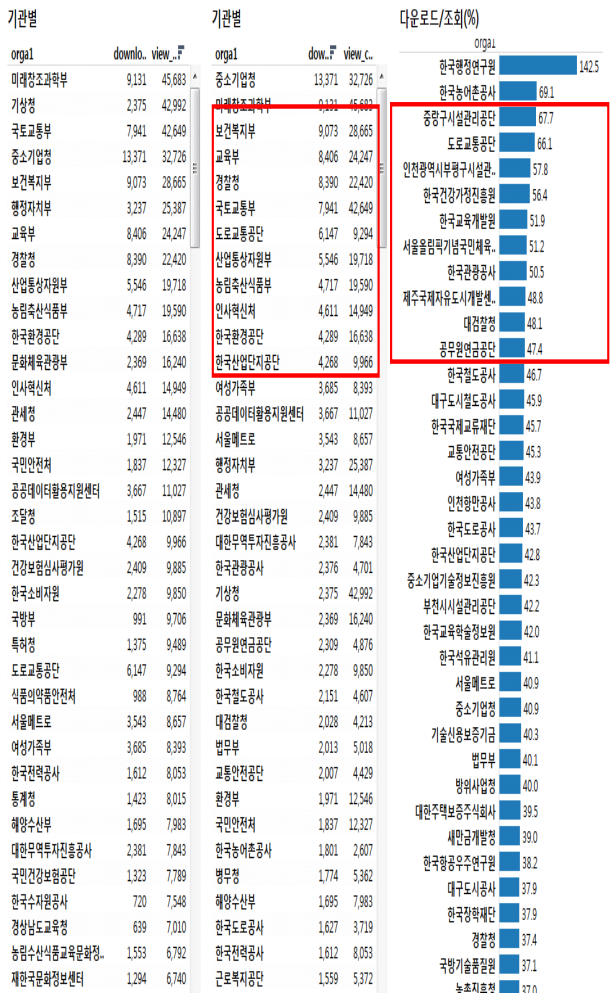


〈그림 11〉 공공데이터 개방 평가지표 2번 결과

2.5.3 개방된 공공데이터 활용 여부 현황

먼저, 전체 기관별 데이터 조회수와 다운로드 수를 비교하면 <그림 12>와 같다.

조회수를 기준으로 한 기관별 현황 결과는 미래창조과학부 데이터 조회수가 가장 많았으며 그 이하로 기상청, 국통교통부, 중소기업청, 보건복지부, 행정자치부, 교육부, 경찰청, 산업통상자원부, 농림축산식품부 순으로 사용자의 관심이 많다는 것을 알 수 있었다. 다음으로는 다운로드 수 기준으로 한 기관별 현황 결과를 보면 중소기업청이 가장 많은 다운로드를 했으며, 그 다음으로는 미래창조과학부, 보건복지부, 교육부, 경찰청, 국토교통부 순으로 나타났다. 따라서 실제 데이터 다운로드 수가 많은 경우 데이터 사용에 있어 양호하다는 것을 추측할 수 있다. 마지막으로 조회수 대비 다운로드수가 많은 기관은 한국행정연구원, 한국농어촌공사, 중랑구시설관리공단, 도로교통공단, 인천광역시 부평구시설관리공단, 한국건강가정진흥원등으로 올린만큼 다운로드 수를 받은 사용자가 많은 것으로 보여진다.



〈그림 12〉 공공데이터 개방 평가지표 3번 결과

전체 기관별 개방데이터 평가지표 결과는 <표 11>과 같다.

〈표 11〉 공공데이터 개방 평가지표 전체 결과

순위	평가지표 1	평가지표 2	평가지표 3
1	산업통상자원부	산업통상자원부	중소기업청
2	국토교통부	농림축산식품부	미래창조과학부
3	보건복지부	특허청	보건복지부
4	행정자치부	미래창조과학부	교육부
5	관세청	보건복지부	경찰청

산업통상자원부의 경우 당연히 산업고용 영역에 많은 데이터를 개방하였고 국토교통부의 경우 또한 교통물류, 국토관리영역에 데이터의 분포 편향되어 있었다. 따라서 평가지표 4번의 경우는 지역별 개방데이터 현황에만 반영하였고 기관별 분석에 있어서는 반영하지 않았다. 평가지표에 따라 기관별 순위가 달라지는 것을 파악할 수 있었다. 예를 들어 산업통상자원부의 경우 데이터 수가 가장 많고, 데이터유형별 평가에서도 3유형 데이터가 많은 기관으로 나타났다 하지만 데이터 다운로드에 있어서는 8위에 머물렀다. 또한 국토 교통부의 경우 데이터수는 많으나 2유형의 데이터 형식이 많아 평가지표 2에서는 Top10 안에 들지 못했다.

Ⅲ. 결 론

본 논문에서는 2013년 공공데이터법률 시행 2년이 지난시점, 공공데이터 개방 평가지표를 개발하여 공공데이터포털 웹에 게시된 공공데이터를 각 지역별, 기관별 개방 현황을 도출하였다. 양적 개방 확대에 비해 산업 활용도는 아직 저조한 수준이다. 제공기관이 선정한 개별 데이터 셋(dataset) 중심으로 개방되어, 산업적 활용성 높은 핵심・대규모 데이터의 개방・활용 미흡하였다. 또한 개방 데이터에 대한 낮은 접근성이 낮았다. 표준화된 메타 데이터 부재 등으로 인해 공공데이터 편리한 검색 및 활용에 한계가 있다 따라서 데이터 품질에 대한 지속적 관리와 평가가 필요하며 향후 연구에서는 자동화 시스템 구축통한 서비스 개선이 필요하다.

참 고 문 헌

[1] 공공데이터 전략위원회, “공공데이터 개방 발전전략(안)”, 2014.

[2] 김지혜, “공공데이터 포털 분석을 통한 서비스 수준 평가”, 졸업논문지, 2016.

[3] 한국지역정보개발원, “공공데이터 제공 방안 및 가이드라인 연구”, 연구보고서, 2014.

[4] 행정자치부, “공공데이터 개방 표준 가이드라인”, NIA한국정보화진흥원, 2015.

[5] 행정자치부, “공공데이터 개방 활용추진 현황”, NIA 한국정보화진흥원, 2015.

[6] http://5stardata.info/.

[7] http://global.census.okfn.org/.

저 자 소 개

김 지 혜(Ji-Hye Kim)

・2014년~현재 : 충북대학교 대학원 경영정보학 (석사)

・관심분야 : 빅데이터 활용

조 상 우(Sang-Woo Cho)

・2014년~현재 : 충북대학교 비즈니스데이터 융합학과 (석사)

・관심분야 : 빅데이터, 데이터마이닝

이 경 희(Kyung-hee Lee)

・2004년 : 충북대학교 컴퓨터과학 (박사)

・2008년~현재 : 충북대학교 비즈니스데이터융합과 연구원 (박사)

・관심분야 : 빅데이터, 알고리즘, 데이터마이닝, 컴퓨팅 플랫폼

조 완 섭(Wan-Sup Cho)

・1987년 : 한국과학기술원 컴퓨터과학과 (박사)

・1996년~현재 : 충북대학교 경영정보학과 (교수)

・관심분야 : 빅데이터, 비즈니스 인텔리전스, ERP