특집 07

오픈 데이터 플랫폼 동향

정유철 · 서동준 · 이혜진 · 김광영 (한국과학기술정보연구원)

목 차

요 약

국/내외의 공공 데이터 공유·개방 흐름에 힘입어, 데이터기반의 다양한 비즈니스 기회가 창출되면서, 데이터를 효과적으로 공유·관리하기 위한 오픈 데이터플랫폼이 공공, 과학기술 분야를 중심으로 확산·발전하고 있다.

공공분야에서는 공공데이터 공유를 위한 CKAN, Socrata 등의 플랫폼이 있으며, 연구분야에서는 DSpace를 기관 데이터 공유 레파지토리(repositories)들이 있다. 국내외에 이러한 플랫폼을 이용하여 데이터를 공유하거나, 분야별로 데이터 저장소들이 증가일로에 있다.

나아가, 최근 단순히 공유하는 것을 뛰어넘어 사용자들에게 데이터 분석을 용이하게 하는 분석·개발·서비스환경을 제공하는 시도가 MS, Google, AWS등에서 보이고 있다. 본 논문에서는 이러한 일련의 플랫폼 개발 동향 및 그들의 특징을 살펴보고, 현존하는 분석형 데이터 플랫폼이 지향하는 기능들에 대해 살펴보기로 한다.

1. 서 론

국/내외의 공공 데이터 공유·개방 흐름에 힘입어, 데이터기반의 다양한 비즈니스 기회가 창출되면서, 데이터를 효과적으로 공유/관리하기 위한 오픈 데이터플랫폼이 공공, 과학기술 분야를 중심으로 확산·발전하고 있다.

또한, 과학계에서 연구 자료를 공개, 공유하는 개방적 연구규범인 '오픈 사이언스'의 흐름에 따라 과학자, 연구자들 간의 지식과 정보를 개방 및 공유하고 협업하는 과정이 웹을 통에 활성화

되었고, 연구 전 과정에서의 연구 개방을 가능케 하기 위한 도구로서의 ICT, 플랫폼 및 인프라 기술들이 결합된 오픈 데이터 플랫폼 구축에 대한 다양한 시도가 이루어지고 있다.

공공분야에서는 공공데이터 공유를 위한 CKAN[1], Socrata[2] 등의 플랫폼이 있으며 북아메리카, 유럽, 아시아, 아프리카대륙의 여러 정부기관에서 데이터공유 포털 구현에 사용하고 있다. 연구 분야에서는 DSpace[3]와 같은 기관 데이터 저장소(repositories)들이 개발되어 전세계 1000+개 대학교, 고등 교육기관, 연구조직 등에서 각자의 기관 저장소구축에 활용되었다.

나아가, 최근 단순히 공유하는 것을 뛰어넘어 사용자들에게 데이터 분석을 용이하게 하는 분석·개발·서비스환경을 제공하는 시도가 MS, Google, AWS등에서 보이고 있다. 본 고에서는 이러한 일련의 오픈데이터 플랫폼 개발 동향 및 그들의 특징을 살펴보고, 분석형 데이터 서비스/플랫폼의 사례 및 분석 기능들에 대해 정리 한다. 그리고, 국내 공공 데이터 개방, 오픈 데이터의 한계, 오픈 데이터 플랫폼 보급현황, 분석형 플랫폼등의 방향성에 대해 간략하게 다룬다.

2. 정부 주도의 데이터 공개 패러다임

Open Data Barometer [4, 5]는 데이터 아젠다별 분석, 국가 및 지역별 분석과 함께 각국의 준비도 (Readiness), 실행력 (Implementation), 효과 (Impact)등의 3개 항목에 대한 국가별 순위를 제공하고 있는데, 각국 정부의 오픈 데이터 정책 추진 방향과 발전 단계 모델을 제시하고 있다. 영국과 미국이 부동의 1, 2위를 점하고 있으며, 오픈 데이터 플랫폼의 발전 양상도 이와 맥락을 같이 한다.

미국은 2012년 다양한 부처가 참여하는 2억 달러 규모의 “빅데이터 연구개발 이니셔티브 (Big Data R&D Initiative)” 내에서 빅데이터 핵심 기술 확보, 사회 각 영역에 활용, 인력 양성의 3가지 측면들을 골자로 하는 세부 계획을 도출하여 실행하였으며, 오픈 데이터 플랫폼인 data.gov를 통해 데이터를 공개하고 있다 [6]. 플랫폼을 통해 공개되는 데이터의 범주는 국방, 지질, 보건, 과학, 공학, 에너지, 조세, 교통 등의 다양한 분야를 포괄하며, 정부 기관 내 데이터를 통합·공유하여 주민 서비스를 개선할 뿐만 아니라, 상당 부분은 민간에서 활용토록 하여 새로운 비즈니스 기회를 제공하였다.

영국의 경우 2012년 “데이터 전략위원회(Data Strategy Board)”를 설립하고, 각 부처에 맞는 오픈 데이터 전략을 발표하였다. 이에 각 부처는 데이터 공유 플랫폼 (data.gov.uk) [7]의 재정비를 통해 데이터 접근성 강화 및 서비스 활성화 방안을 모색하고, 오픈 데이터 평가 방법을 도입하게 된다 [6]. 영국의 경우 데이터를 공개하는데 그치지 않고 data.gov.uk의 개발 코드를 오픈 소스인 CKAN (ckan.org)을 통해 관련 오픈 소스로 공개하고 있다

미국과 영국의 사례에서 볼 때, 정부 주도의 데이터 공개는 우선 공공부문의 패러다임 변화를 의미하며, 국가사회 현안 해결을 위해 복지, 의료, 재난, 교통 등의 다양한 데이터를 활용하려는 노력을 확인할 수 있다. 해외 대다수의 선진국들에서 오픈 데이터 프로그램을 실시하고 있지만, 아프리카 대륙의 케냐의 사례 [8]는 매우 모범적이다. 농업, 교육, 에너지, 환경, 재정, 건강, 물과 위생 등의 11개 분야의 데이터를 적극적으로 개방하고 민간에서의 활용을 촉진하기 위한 노력을 보이고 있다. 케냐의 낮은 인터넷 보급률 및 모바일 접근도를 고려할 때, 다른 선진국 보다 더 적극적인 개방형 혁신사례라 볼 수 있다.

우리나라 정부는 공공정보 개방 패러다임에 따라 실수요자 중심의 맞춤형 정보 개방 계획을 수립하고, 다량의 공공정보를 개방하고 있는데 [9], 대표적인 데이터 포털들로는 한국정보화진흥원에서 운영하고 있는 공공데이터포털 [10] 및 서울시의 열린 데이터 광장[11]이 있다. 이 포털들은 국가가 보유하고 있는 다양한 공공정보를 개방하고, 서비스 유형 (데이터 또는 Open API), 제공기관, 분류체계, 이용허락범위, 태그, 확장자에 따른 다양한 검색기능을 제공하여 국민들이 손쉽게 활용할 수 있게끔 지원하고 있다. Data.go.kr의 경우 그 개방의 규모가 현재(2016년 7월말) 개방기관 658개, 파일데이터 14,826, 오픈API 1,990개에 이르고 있다.

하지만 국내 공공 데이터 포털들은 단기간에 상당한 수의 데이터를 확보하였으나, 데이터의 제공형태가 제한적이며 미국/영국의 공유 데이터 플랫폼과 비교할 때 산업적 활용과 서비스 다양성은 미흡하다. 효과적인 활용을 위해서는 공공데이터의 품질을 제고하고, 관련 ·법제도를 개선하여 공공데이터의 공유 문화가 조성될 수 있는 기반을 구축함과 동시에, 상호 연계, 변환, 검색, 시각화 할 수 있는 플랫폼 기술의 개발에도 꾸준한 노력을 기울여야 한다.

더불어서, 정부 주도로 개방되고 있는 공공데이터들은 정책의 중심이 개방에서 다른 곳으로 옮겨지면 관리가 미흡하게 되고 심지어는 소실이 발생할 수 있다. 그래서 공공데이터에 대해 정리 표준화를 거쳐 장기적인 접근이 가능하도록 공공의 저장소에 보관하고, 정보환경이 변하더라도 다음 세대가 같은 정보량을 이용할 수 있도록 지금 세대의 적극적인 노력이 요구된다. 특히, 과학기술분야의 경우 정부출연연구소들에서 국가 연구개발 사업으로 생산되는 연구데이터는 논문, 특허 등으로 발표된 후라도 원시데이터를 연구자의 권한으로 공개, 공유하고, 이를 재사용한 연구자는 인용으로 보상하는 문화의 확산이 필요하다.

3. 오픈 사이언스 시대의 플랫폼의 역할 및 동향

과학계에서 부는 바람인 '오픈사이언스'는 과학계에서 연구 자료를 공개, 공유하는 개방적 연구규범을 지칭한다. 사실 '오픈 사이언스'는 고대 과학계에서부터 전해져 내려오는 과학자들의 과학적 깨달음의 공유에서부터 시작된 것이며, 21세기 들어서 IT 기술의 발달로 인해 다양한 사람들이 참여 가능한 '개방형 과학'으로 확산되어 온 것이다. 2004년부터 OECD를 중심으로 오픈 사이언스 프로젝트가 구체화 되고 있으며, 공공연구 성과물(출판물 및 데이터)에 대한 접근성을 제고하는 차원에서 오픈 사이언스를 논의해 왔다.

2015년 OECD를 중심으로 오픈 사이언스에 대한 범주를 '오픈 엑세스', '오픈 데이터', 그리고 이를 가능케 하는 기술을 포함하는 '오픈 콜라보레이션'으로 정의하였다[12].

오픈 콜라보레이션은 연구 전 과정에서의 오픈 엑세스, 오픈 데이터를 가능하게 하기위한 도구로서의 ICT 및 플랫폼기술의 응용 등을 의미하며 연구의 개방을 위한 협업을 가능하게 하는 연구 인프라까지 포함한다.

이러한 측면에서 보다 효율적이고 활용가능한 오픈 데이터 플랫폼 구축에 대한 시도가 이루어지고 있다.

4. 기관 데이터 공유 플랫폼 – DSpace

연구자들의 연구 데이터를 공유하기 위해 개발되어 널리 쓰이고 있는 플랫폼으로 DSpace

[3]가 있다. DSpace는 2002년 MIT 대학과 HP 연구소의 개발자들의 협력에 의해 공동으로 개발하여 무료로 공개하고 있는 소프트웨어로써 현재 1000+개 이상의 대학교, 고등 교육기관, 문화조직 임 및 연구센터에서 사용되고 있다. 최초 MIT의 논문, 회의자료, 이미지, 동료간의 리뷰 자료, 기술보고서, 연구중인 자료 등에 대한 저장, 공유, 검색을 위한 기능들이 개발되었으며 꾸준히 안정화 및 추가 수정·개발되어 2016년 현재 5.5버전에 이르고 있으며 그 소스 코드는 GitHub (https://github. com/DSpace/DSpace)를 통해 공유 되고 있다.

DSpace에서 구축 가능한 자료의 범위는 학술 논문, 책, 회의보고서, 3D, 사진, 필름, 비디오, 연구데이터를 포괄한다. 최초 교수 연구자료만을 수집하였으나, 점차적으로 유용성이 인정되는 학생 및 기타 연구원들의 자료들 (학위 논문 및 다양한 지적 자산들)에 대해 저작권이 허락하는 한 보관할 수 있게 하였다.

이렇게 다양한 연구 데이터들을 관리, 저장, 검색을 위해 정의된 4가지 기능들은 다음과 같다.

(1) 데이터 관리 기능: 기관 레파지토리의 성격에 맞게 조직 구조를 반영할 수 있는 데이터 모델을 설계한다. 그리고 콘텐츠의 기술, 관리 구조를 위한 3가지 메타 데이터로 구성된다. 다양한 데이터의 형식

(2) 사용자 및 권한 관리: 이용자 식별을 위한 사용자 및 그룹 관리 기능이 제공되며, 인증방식을 선택적으로 설정할 수 있게끔 스택 형식의 인증구조를 따르고 있다. 사용자 그룹별 자원 객체의 접근권한 정책을 다르게 설정 할 수 있다.

(3) 저장관리: 비트 스트림에도 영구적 식별자를 부여하며, 콘텐츠 저장, 관리를 위한 견고한 스토리지 자원 브로커를 지원한다. 또한 콘텐츠가 파손되었는지 검사하는 무결성 체크 기능도 제공한다.

(4) 검색/열람: 널리 알려진 아파치 루씬 (Apache Lucene)[13]기반의 검색서비스와 다양한 브라우징 기능을 제공한다.

Dspace이외에도 Fedora [14], Eprints[15] 등의 연구데이터 저장소 플랫폼이 널리 사용되고 있다. 이들의 최근 움직임은 상업적 정보 커뮤니티와 학술적 커뮤니티의 특성을 변화시키고 있다. 대학, 연구소, 출판사, 도서관, 기업 등은 혁신적인 리포지터리 기반 시스템을 생성하고 있다. 이 시스템들은 디지털 컨텐츠의 생성과 관리를 지원하는 기본 기능에서부터 정보의 생애 전체를 고려한 정보의 이용, 재이용, 상호연결 및 장기보존과 아카이빙등의 다양한 부가적인 기능을 지원하고 있다.

5. 데이터 공유 플랫폼들

정부·공공기관이 보유한 데이터를 공유하기 위한 플랫폼으로는 대표적으로 CKAN[1]과 Socrata[2]가 있다. CKAN은 비영리 단체인 Open Knowledge Foundation (OKF)에 의해 개발되었으나 영국, 미국, 캐나나 등 40개 이상 국가에서 활용 중이며, 기본 기능 이외의 시각화나 API추출등의 특화기능들은 Drupal [16]과 같은 타 오픈 소스와 결합하여 발전시키고 있다. CKAN은 자체적으로 제공하는 플랫폼의 기능을 사용할 수도 있으며, CKAN API만을 가지고 별도의 서비스 제공이 가능하다.

Drupal은 개인 또는 커뮤니티가 웹사이트의 다양한 자료들을 손쉽게 관리, 조직, 출판할 수 있도록 다양한 기능을 제공하는 오픈소스기반의 컨텐츠 관리 시스템으로 미 정부를 중심으로 많은 기관들이 사용하고 있다. 오픈소스, 빠른 개발, 저비용, 확장성, 모듈형 구조, 싱글설치 멀티

유즈, 접속성을 강점으로 세계의 수많은 사용자들과 기관들로 배포하고 있다.

Socrata는 미국 정부의 data.gov 및 뉴욕과 시카고를 포함한 10여개 이상의 주정부 데이터 포털에서 사용 중인 데이터 공유플랫폼이다. 클라우드를 기반으로 기관 보유 데이터의 표준관리, 신규 생성 데이터 자동관리, 데이터 제공자 및 이용자 접근 가능한 API 제공 한다. 오픈 소스인 CKAN에 비해 시각화 및 분석기능이 강점이다.

데이터 공유 플랫폼은 기본적으로 데이터를 편리하고 손쉽게 개방·활용하는 기반으로써 주로 다음과 같은 5가지 공통기능을 갖고 있다.

데이터 공유플랫폼에서 제공할 수 있는 데이터의 형태는 데이터 원본, 데이터 셋, API 등의 프로그램들이다. 이들 데이터는 주로 관리자에 의해 업로드 되며, 관리자는 정부 및 공공기관 등록을 통해 권한을 부여 받는다.

데이터 공유 플랫폼의 기본 관리기능들을 이용하여, 체계적으로 데이터를 관리하는 것이 핵심이며, 특히 입수 및 발행 프로세스의 표준화, 효과적인 공유를 위한 표준화된 메타데이터의 개발, 활발한 오픈 소스 활동을 통해 민간/공공에 적은 비용으로 플랫폼을 공유케 하는 것이 중요하다. 오픈소스 기반의 데이터 플랫폼은 여러 시행착오를 감수해야 하지만, 무료이기에 데이터 이용 활성화 측면에서 바람직한 방법이라 할 수 있다. 최근 미국의 data.gov 포털도 CKAN을 도입하는 추세이다.

〈표 1〉 오픈 데이터 플랫폼의 공통기능 ([6]에서 발췌)

주요 기능	기능 설명
데이터 등록	· 데이터 개방을 위해 데이터 업로드, 데이터 저장등의 역할을 수행 · 데이터 등록을 위해 지원가능 한 파일 형식을 정의하고, 데이터의 API 추출 등의 기능 제공
데이터 발행	· 등록된 데이터에 대해 지원 가능한 파일형태로 데이터를 제공 · 데이터 제공에 따른 보안, 표준등 관리 기능
데이터 현황	· 사용자 및 관리자를 위해 데이터 출처, 분류, 내용 등의 포함된 데이터 카탈로그 기능 · 데이터의 원활한 사용을 위한 다양한 검색 기능
데이터 포털	· 사용자, 고객센터, 포털 운영 관련 서비스 지원 기능
시각화 기능	· 플랫폼 기술 수준에 따라 데이터 추출 및 필터링을 통한 도표, 그래프 등 시각화 기능 등

플랫폼을 사용하는 것은 단순히 데이터를 활용에서 시작하여 플랫폼을 기반으로 축적된 지식과 경험을 공유하여 사회현안 해결 및 새로운 비즈니스 창출을 위해 민간에서 활용토록 하는 데 의의가 있다. 데이터의 효과적 활용을 위해서는 플랫폼 핵심 기능(생산, 저장, 관리)뿐만 아니라 인프라적인 기능(유통, 활용), 데이터 분석을 위한 핵심기술 개발, 그리고 데이터 활용을 위한 인력개발 등 데이터 분야 전반에 대한 노력이 필요하다.

국내에서는 CKAN 및 DKAN의 일부 기능을 Data.go.kr 포털에 반영하는 수준에서 데이터 공유 플랫폼을 이용하고 있으며, 해외와 같이 적극적인 개발 플랫폼 공유 시도는 미미하다. 다만 2015년부터 빅데이터 관련 기업들이 한데 모여 오픈데이터플랫폼 (ODP) 이니셔티브란 연합체를 출범시켰으며, 엔터프라이즈에서 쉽게 사용 가능한 표준 빅데이터 플랫폼 개발을 목표로 협력을 강화하고 있다. 한국 오픈데이터플랫폼 이니셔티브는 벤더 주도의 연합체로써 해외의 데이터 공유 플랫폼 동향과는 차이가 있지만, 그만큼 오픈 데이터를 다루는 문제가 기업의 이윤창출을 위해 중요한 영역으로 오고 있다는 것을 반증하는 것이라 볼 수 있다.

6. 분석형 데이터 서비스/플랫폼

6.1 데이터 기반 분석 서비스/플랫폼의 사례

국내외 사례에서 데이터 분석 플랫폼은 주로 교통, 화재, 지진과 같이 공공의 사회문제 해결을 위해 도메인 특화 플랫폼을 개발한 경우와 공공 논문·특허와 같은 지식자원을 분석 가공하여 서비스를 제공하는 사례들이 있다.

6.1.1 사례1: 교통 데이터 기반의 분석형 서비스

INRIX사 [17] 는 교통관련 분석 정보 서비스를 제공하는 분석형 데이터 플랫폼을 운용하고 있다. 수년간 수집된 교통 데이터를 기반으로 지도, 실시간 교통정보, 네비게이션과 같은 서비스를 기본으로 제공하고 있으며, 공공 분야에서는 교통예측, 유료도로 통행료 측정, 실시간 교통정보 서비스 등을 제공하고 있다. 뿐만 아니라 교통 정보의 플랫폼으로써 모바일 개발자를 위한 솔루션 또한 제공하고 있다.

6.1.2 사례2: 지진 예측 분석 플랫폼

지진 관련 재난 데이터의 패턴을 분석하고 예측한 현재의 상황과 사물인터넷(IoT) 센서 데이터를 통해 수집된 현재의 데이터와 비교하여 전조현상으로 감지되는 지진 발생 예측의 정확도를 높이는 시도 [18]가 있다. 물론 정확하고 빠른 처리를 위해서는 많은 양의 데이터를 빠르게 분산처리해야하는 과정이 필수적이다.

6.1.3 사례3: 논문 · 특허 기반 분석 서비스

한국과학기술정보연구원(KISTI)는 해외 특허, 논문의 지식자원과 기업 및 제품 정보를 가공·분석하여 중소기업 지원을 위한 분석 서비스들을 2015년부터 개방해오고 있다. 플랫폼 기업 보유 제품을 기반으로 새로운 기회제품 영역을 탐색하고, 관심기업현황도 분석할 수 있는 기술기회 탐색시스템 TOD (Technology Opportunity Discovery) 서비스 [19] 와 글로벌 경쟁기술의 활동상황을 분석할 수 있는 경쟁정보분석시스템 COMPAS (Competitive Analysis Service) [20] 이 있다. TOD와 COMPAS는 신사업 기회를 탐색하고, 경쟁기술 및 경쟁기업 정보를 지속적으로 모니터링하기 원하는 국가연구개발사업 연구자 및 기술혁신형 기업을 타겟 사용자로 하고 있다.

TOD는 크게 세 가지 분석 서비스로 구성되어 있다. (1) 이용자가 보유중인 제품 및 기술과 관련 있는 기회제품을 추천해주는 '보유제품기반 기회제품 탐색서비스' (2) 기업 간 제품별 경쟁현황을 비교 · 분석해 기회제품을 추천해주는 '경쟁기업 벤치마킹서비스' (3) 제품의 기능 및 기술을 탐색해 적용 가능한 분야 및 기회제품을 추천해주는 '제품 · 기술 관계 분석서비스' 등이다.

COMPAS는 기존에 서비스됐던 핵심 경쟁자 탐색, 핵심 경쟁자 프로파일, 유사특허 탐색, 무역역조 탐색, 인용트리 탐색 모듈에 이어 핵심특허 탐색, 주목 특허기술 탐색, 테크트리, 기술경로 탐색 등의 분석모듈들을 제공한다.

6.2 분석형 데이터 플랫폼의 기능들

데이터 분석 측면에서 볼 때, 기본적으로 대상 데이터를 간단하게 보기 위한 필터링과 및 시각화기능을 시작으로 사용자 주도의 데이터 가공-기계학습-분석형 웹서비스에 이르는 심층적 분석 기능으로 나눠볼 수 있다.

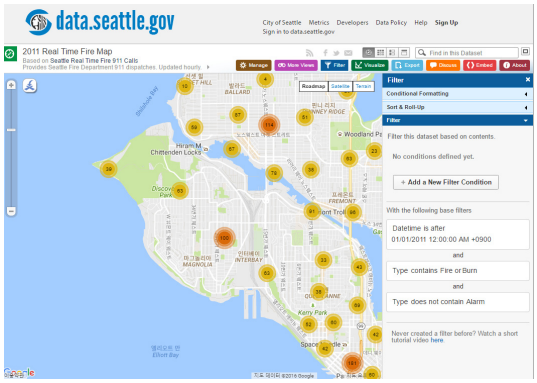
6.2.1 엑셀 수준의 간단한 필터링 및 시각화 기능

플랫폼내에 파일이 적재되는 경우, 특정 필드를 지정하고 기준 값을 설정하여 해당 조건을 만족하는 특정 데이터을 필터링하고 이를 시각화 할 수 있다. 미국의 경우, 시각화에 매우 유용한 Socrata API를 활용하여 데이터 시각화 서비스를 제공 중이며, 구글 맵 등의 서비스와 매쉬업 하여 데이터와 콘텐츠의 융합 서비스를 제공하고 있다. 미국 최초 911의 실시간 화재 및 범죄 데이터, 건축 허가정보 등 370여개 데이터 셋을 Socrata API(SODA API) [21]를 통해 가시화 서비스, 데이터의 export 기능 들을 제공 하고 있다. 아래의 예는, 시애틀 소방국에 의해 관리되고 화재정보이며 5분마다 업데이트 되어 공개되고 있는 오픈 API 및 그것의 가시화 결과이다.

· SODA API 호출 예시

(https://data.seattle.gov/Public-Safety/Seattle-Real-Time-Fire-911-Calls/kzjm-xkqj)

구글 퓨전 테이블(Fusion Tables)[22]은 클라우드 기반의 스프레드시트(spreadsheet)로써, 단순히 데이터를 입력하고 관리하는 기능뿐만 아니라 자신이 가지고 있는 데이터를 이용하여 시각화하는 기능을 통해 지도와도 연계할 수 있다. 지도 시각화로 연결될 수 있는 데이터가 주소와 같은 위치 참조 속성을 가질 경우, 퓨전테이블의 지오코드 기능이 이를 자동으로 지리적 좌표와 연결해주고, 밀도맵, 점묘도 등으로 데이터의 공간적 분포를 분석할 수 있도록 지원한다.



〈그림 1〉 화재정보의 시각화 결과 (data.seattle.gov)

6.2.2 클라우드 기반의 데이터 분석 및 서비스 기능

마이크로소프트, 구글 등의 기업에서는 데이터를 사용자의 목적에 맞게 분석, 공유, 웹서비스화 할 수 있는 클라우드 환경의 플랫폼을 개발하여 웹으로 서비스하고 있다.

a) 마이크로소프트의 클라우기반 데이터 플랫폼

MS의 Azure 플랫폼[23]은 데이터를 손쉽게 조작할 수 있는 프로그래밍 환경을 제공하거니와, 데이터 변환 및 대화형 데이터 시각화 툴, 기계학습, 빅데이터 사용을 용이케 하는 분산 분석 서비스, 대규모 저장소, 기계학습 인프라를 제공한다. 데이터 학습기반의 분류 및 예측 서비스(예, 감성분류)를 바로 웹서비스로 연결할 수 있는 편리한 개발·서비스 환경을 클라우드로 제공한다.

b) 구글의 클라우드기반 데이터 플랫폼

구글 빅쿼리 (BigQuery) [24]는 대용량 데이터(최대 몇 십억 개의 행)를 UI기반의 대화형 인터페이스로 접근하여 분석할 수 있는 플랫폼으로써, SQL과 유사한 질의로 대규모 데이터를 핸들링 할 수 있다. 클라우드 기반으로 별도의 인프라 없이 대용량의 데이터를 빠르게 추출, 가공, 분류 및 분석결과를 제공하는 솔루션이다.

구글의 Cloud Datalab [25]은 대규모의 데이

터를 탐색, 분석, 시각화 할 수 있는 대화형 데이터 툴이다. 기 존재하는 구글의 클라우드 플랫폼들인 빅쿼리(BigQuery), 컴퓨트엔진(Compute Engine), 그리고 클라우드 스토리지(Cloud Storage)와 결합하여 사용될 수 있으며 Git기반의 공동 개발환경을 제공한다. 또한 python, sql, javascript등의 다양한 개발언어를 지원하고 있다.

c) KISTI의 클라우드 기반 데이터 플랫폼

KISTI에서는 2015년부터 3개년 계획으로 S&T 지식플랫폼을 설계·구축하고 있으며, 2016년 말 KISTI 내부 핵심 과학 컨텐츠 (i.e., 논문, 특허, 연구보고서)를 탑재한 데이터 공유 플랫폼 공개를 앞두고 있다. 관리자가 아닌 인증된 연구자들이 자유로이 본인의 데이터를 공유할 수 있으며, 표준화된 입수/공개 프로세스에 의해 DOI (Document Object Identifier)까지 할당하여 데이터를 공유케 한다. 이외에도 연구자 활동에 도움이 되는 다양한 데이터 셋, 오픈 API, 데이터 활용 SW를 공유 및 활용할 수 있는 환경을 제공한다.

향후에는 KISTI가 개발한 대규모 병렬 Matrix 계산 플랫폼[26]과 결합하여 대규모의 연구데이터에 대한 계산환경을 클라우드체제로 전환할 예정이다. Python 수준의 script 프로그래밍이 가능한 연구자라면 이 플랫폼에 접근하여 유수의 해외 상용 솔루션 (e.g., MATLAB, SPSS, 그리고 Amazon Web Service (AWS) 보다 빠른 계산·분석 결과를 획득할 수 있다. 과학기술 연구자의 연구활동을 전방위로 지원하기 위해 국내에서는 최초로 개발되는 연구자 지원 분석형 플랫폼이라 할 수 있다.

기존의 오픈 데이터 플랫폼이 데이터의 공유·개방에 가치를 두었다면, 분석형 데이터 플랫폼은 보다 다양한 연계·활용에 중점을 두고 있다고 볼 수 있다. 특정 도메인에서 사용자들의 편익을 위한 분석형 서비스를 기획하거나, 도메인을 고려하지 않고, 연구자 또는 개발자가 클라우드 환경에서 자유로이 대량의 데이터를 분석하고, 목적에 맞는 학습 알고리즘 고려한 예측/분류 모듈을 구현하며, 이에 기반한 서비스를 개발할 수 있는 환경을 제공하는 것이다. 앞으로 기 공개된 공공데이터 및 오픈 API를 손쉽게 적재 (import)하여 분석형 서비스로 만드는 다양한 사례가 소개될 것으로 기대된다.

7. 토 론

국내 공공 데이터 개방 수준: Open Data Barometer (ODB) 모델에서는 정부의 정책수립과 정책 추진 의지를 토대로 결국 얼마나 적극적으로 실행하느냐가 국가의 ODB 점수를 결정하게 된다. 한국은 2013년 총 77개국 대상 조사에서 12위를 차지 한 후, 2015년에는 총 92개국 대상 조사에서 9위를 차지한바 있다 [4, 5].

오픈데이터의 한계: 세계적인 오픈 정부 데이터 계획 및 정책에 힘입어, 기존에 없던 공공 데이터가 공개되었지만, 거의 90%는 아직 닫혀있다 [5]. 또한 공개된 10% 역시 데이터 품질이 형편없으며, 효과적으로 접근, 활용하기는 매우 어려운 것이 현실이다.

국내 상황도 크게 다르지 않다. 공개된 데이터가 대부분 업무용 DB이기도 하거니와 민간이 바로 활용하기에는 오류율이 높다. 또한 충분한 활용 예시가 미미하여, 데이터를 정제하고 연결하여 사용하기 위해서는 상당한 시간·비용이 수반된다. 사실 데이터를 보다 잘 처리하여 다각적인 분석에 활용할 수 있는 가능성은 열려있으나, 데이터가 체계적으로 표준화되어 있지 않아, 그 활

용이 제약적이다.

오픈 데이터 플랫폼 보급 현황: 데이터 보유자와 이용자를 상호 연결해 주는 서비스로 데이터 개방을 위한 기회와 위험요인을 모두 내포한다. 오픈소스 기반 데이터 플랫폼의 사용자의 참여를 통해 데이터 활용성을 극대화할 수 있는 효과적인 반면, 상용 솔루션에 비해 관리자의 부단한 노력 및 시행착오를 요구할 수 있다. 미국의 사례에서 볼 때 최초에는 소크라타사의 플랫폼을 기반으로 data.gov 사이트를 개발하였으나 오픈 소스 플랫폼인 CKAN의 확산에 따라 점진적으로 CKAN 또는 Drupal기반의 DKAN [27]을 적극적으로 도입하고 있다. 현재는 상용 플랫폼과 오픈 소스 플랫폼의 사용이 공존하는 단계이나, github와 같은 공동 개발 환경에서의 협력이 계속 활성화된다면 오픈 소스기반 플랫폼의 사용이 더욱 확대될 것이라 사료된다.

분석형 플랫폼의 방향성: 공공의 이익 증진 및 사회 현안 해결을 위해 특정 도메인에서의 분석 플랫폼이 구축되어 널리 활용되고 있다. 과거 데이터 축적이 가능한 금융, 검색, 마케팅 분야의 스타트업이 시장을 주도하였으며, 최근 데이터 분석 자체를 기반으로 하는 다양한 스타트업이 등장하는 흐름을 볼 때, 분석기반의 데이터 플랫폼은 더욱 그 중요성이 더해 질것이며, 효과적 데이터의 연계·융합·활용 문제에 대한 솔루션이 보다 다양하게 개발 될 것을 추정된다.

8. 결 언

정부 주도의 공공 데이터 공개 패러다임은 세계적인 추세이며, 주요 몇 개 국가에서 공개된 데이터의 다양성과 개수는 이미 상당한 수준이다. 다만 개인 또는 민간 기업에서 바로 활용할 수 있는 고품질의 효용성 있는 데이터의 공개가 관건이라고 하겠다.

과학 분야에서의 오픈사이언스(Open Science) 운동은 연구논문, 연구 데이터등의 공유를 지칭하는 Open Access, Open Data등의 슬로건으로 전 세계적으로 확대되고 있는데, 최근 유럽연합(EU)가 2020년부터 누구나 과학논문을 무료로 열람케하는 시도와 큰맥을 같이 한다.

이러한 세계 각국 정부 및 과학기술계의 '데이터 개방', '지식의 개방화' 는 사용자들이 손쉽게 참여하여 공유할 수 있는 기반이 되는 오픈 플랫폼의 개발을 촉발 시켰다. 최근 MS, Google, AWS의 상용 클라우드 플랫폼은 대량의 데이터를 손쉽게 조작, 분석, 기계학습하고, 플랫폼 내에서 웹서비스로 연결하는 통합 솔류션을 제공하고 있다.

국내외 공유·개방의 흐름을 꾸준히 모니터링하고, 오픈데이터 플랫폼에서 활용되는 융·복합 기반기술 및 SW기술에 대한 꾸준한 연구와 함께, 공유데이터의 효율적 활용을 통한 새로운 비즈니스 기회를 창출의 기회 및 환경을 조성하는데 꾸준한 노력을 기울여야 할 것 이다.

참 고 문 헌

[1] CKAN, The open source data portal software, http://ckan.org/

[2] Socrata, The open data platform for digital government, https://socrata.com/

[3] DSpace, http://www.dspace.org/

[4] Open Data Barometer - 2013 Global Report (Open Data Barometer, 2013), Available at: http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf

[5] Open Data Barometer - 2015 Global Report (Open Data Barometer, 2015), Available at: http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf
[6] 오픈데이터 플랫폼과 국가 데이터 전략방향, IT & Future Strategy, 제16호 2013.12
[7] 영국 데이터 공유 플랫폼, http://data.gov.uk
[8] 케냐 오픈 데이터 포털, https://opendata.go.ke/
[9] 송인국, 인터넷기반 공공데이터 활용방안 연구, 인터넷정보학회논문지, 16(4), pp. 131-139, 2015.
[10] 대한민국 공공 데이터 포털, http://data.go.kr
[11] 서울시 열린데이터 광장, http://data.seoul.go.kr
[12] OECD (2015), "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en
[13] Apache Lucene 검색엔진, https://lucene.apache.org/core/
[14] Fedora Repository, http://fedorarepository.org/
[15] Eprints, https://en.wikipedia.org/wiki/EPrints
[16] Drupal, A Open Source CMS, https://www.drupal.org/
[17] INRIX사 홈페이지, http://inrix.com/
[18] IoT와 재난 데이터를 활용한 지진 예측 분석 플랫폼
[19] 기술기회탐색 플랫폼 (Technology Opportunity Discovery, TOD), http://tod.kisti.re.kr/index.do
[20] 경쟁분석서비스 (COMPetitive Analysis Service, COMPASS) http://compas.kisti.re.kr/index.jsp
[21] SODA API, https://dev.socrata.com/consumers/getting- started.html
[22] Hector Gonzalez, Alon Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen (2010). Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud, SoCC'10.
[23] MS Azure Platform, https://azure.microsoft.com/
[24] 구글 빅쿼리, https://developers.google.com/bigquery/
[25] 구글 클라우드 데이터랩, https://cloud.google.com/datalab/
[26] Tupix 빅데이터 분석시스템, http://www.dongascience. com/news/view/12135
[27] DKAN, http://docs.getdkan.com/

저 자 약 력

정 유 철

이메일 : jyc77@kisti.re.kr

- 2011년 한국과학기술원 전산학과 (박사)
- 2009년~2013년 한국전자통신연구원 선임연구원
- 2013년~현재 한국과학기술정보연구원 선임연구원

서 동 준

이메일 : djsuh@kisti.re.kr

- 2014년 KAIST 건설및환경공학과 (건설IT융합전공) (박사)
- 2014년~2015년 KAIST IT융합연구소 연구조교수
- 2015년~현재 한국과학기술정보연구원, 선임연구원

이 혜 진

이메일 : hyejin@kisti.re.kr

- 2013년 숙명여자대학교 문헌정보학과 (박사수료)
- 2002년~현재 한국과학기술정보연구원 선임연구원

김 광 영

이메일 : glorykim@kisti.re.kr

- 2011년 충남대학교 문헌정보학과(박사)
- 2001년~현재 한국과학기술정보연구원 선임연구원
- 2015년~현재 한국과학술정보연구원 정보융합연구실장