

## 혁신의 유효숫자

지난 대선부터 본격적으로 화제가 된 4차 산업혁명은 여러 논쟁을 거쳐 이제는 대한민국의 많은 사람에게 보편적으로 받아들여지고 있다. 4차 산업혁명의 이미지는 급진적이고 전례 없는 변화로 그려진다. 세계경제포럼의 클라우드 슈밥은 4차 산업혁명은 인류가 지금까지 경험했던 것과는 전혀 다른 규모, 범위, 복잡성을 몰고 올 것이라고 말한다. 단순히 몇몇 새로운 서비스의 등장이나, 사회 전반의 모습이 바뀔 것이라는 예측이다. 인공지능과 로봇 기술이 발전하여 각종 산업의 모습을 바꾸고, 이에 따라 현존하는 여러 직업이 사라지고 새로운 인재상과 재능이 중요해질 것이라고 한다. 4차 산업혁명의 대표적인 기술로 꼽히는 자율주행차가 보급되면, 극단적으로는 차를 소유하는 개념이 없어지고 필요에 따라 차량을 호출해서 사용할 것이라고 전망하는 사람들도 많다. 소유의 개념까지도 변화하는 것이다. 4차 산업혁명에 대한 이러한 묘사는 마치 계단 오르기를 연상케 한다. 3번째 계단과 4번째 계단 사이의 높이가 갑자기 쑥 올라가는 것처럼, 지금까지 생각하기 어려웠던 놀라운 변화가 사회 곳곳에서 순식간에 벌어진다는 뜻이다.

### 데이터 혁신 촉진 위한 공공데이터 개방과 활용

그중에서도 데이터는 4차 산업혁명의 주요 기반으로 꼽힌다. 사물인터넷은 우리 일상생활 곳곳에서 다양한 데이터를 수집하는 센서로 기능하고, 인공지능은 이 데이터를 학습하여 판단의 근거를 만든다. 데이터를 기반으로 한 혁신을 촉진하기 위해 우리 정부도 여러 방면으

로 노력하고 있다. 4차 산업혁명이란 말이 등장하기 이전부터 공공데이터 개방을 위해 2011년부터 공공데이터포털(www.data.go.kr)을 구축하여 건축, 교통, 건강, 부동산 거래 등 약 25,000건에 이르는 다양한 데이터셋을 제공하고 있다. 공공데이터 이용에 따른 규제도 강하지 않아서, 공공데이터법에 의거하여 영리 목적으로도 자유로이 사용할 수 있다. 정부가 가진 데이터를 혁신역량이 있는 기업과 개인에 개방하여 데이터의 창의적인 이용을 촉진하겠다는 뜻이다.

공공데이터 개방에는 지자체도 적극적으로 참여하고 있다. 서울시는 2013년부터 서울열린데이터광장(data.seoul.go.kr)을 통해 지하철 도착 시각부터 대기오염 실태, 공동주택 현황 등 생활에 밀접한 정보를 상업적 이용을 제한하지 않고 공개하고 있다. 이에 더해 2016년부터는 대표적인 소스코드 저장소인 깃허브(github.com)에 500만 건 이상의 행정정보 목록을 공개하여 더 많은 개발자가 데이터를 활용할 수 있도록 장려하고 있다. 강원도, 경기도, 부산광역시, 대전광역시 등 대부분의 광역지자체가 마찬가지로 각자의 공공데이터 포털을 운영하고 있다.

공공데이터를 둘러싼 목소리는 단순하다. 더 많은 데이터를 공개하라는 것이다. 이들은 대부분의 데이터는 공개하지 않을 이유가 없다면서, 공개할 수 없도록 지정한 자료를 제외한 데이터는 모두 개방하는 네거티브 데이터 정책을 주장한다. 데이터를 어떻게 써야 할지는 혁신역량이 뛰어난 민간이 가장 잘 알고 있기 때문에, 가리지 않고 데이터를 개방해야 한다는 말이다. 더



글\_윤기준 | KAIST 과학기술정책대학원 박사과정 ykjstp@gmail.com

KAIST 건설 및 환경공학과 졸업 후 동대학원에서 과학기술정책으로 석사학위를 받았다.

많은 정보를 공개하면 정부의 투명성을 담보할 뿐만 아니라, 대부분의 정보를 공개 망에서 저장하고 운영하기 때문에 이를 다루는 공무원 업무의 효율성까지 재고할 수 있다는 장점도 주요한 근거다. 데이터만 개방하면 4차 산업혁명이라는 계단을 쑥 하고 오를 것이라는 낙관적인 관점이다.

### 원하는 만큼 활용 쉽지 않은 공공데이터

그러나 데이터를 둘러싼 이슈는 단순히 개방을 하느냐 마느냐의 문제가 아니다. 정작 4차 산업혁명의 역군이라 불리는 사람들 앞에는 작고 사소해 보이지만 중요한 계단이 수없이 펼쳐져 있다. 데이터가 공개되어도 그것을 잘 사용하기 위해서는 생각보다 많은 정제와 검증 작업이 필요하기 때문이다. 예를 들어 전국 도로명주소를 모아둔 데이터베이스가 필요하다고 생각해 보자. 개발자는 이 데이터가 정확히 어떤 이름으로 이 데이터가 시스템에 등록되어 있는지 알지 못한다. 따라서 도로, 도로명, 전국 도로 등으로 검색을 시도할 것이다. 그러나 그 어떤 키워드로 검색을 해도 도로 시설물 현황, 자전거 도로 통계 등 원하지 않는 결과 수백 개가 함께 나와서 개발자는 일일이 스크롤을 내려가며 원하는 데이터를 찾아야 한다. 그나마 원하는 데이터가 나오면 다행이고, 검색 결과에 원하는 데이터가 없는 경우 새로운 키워드를 이용해서 검색을 다시 시도해야 한다. 데이터가 한두 종류이면 큰 지장은 없겠으나, 여러 데이터를 융합하여 새로운 가치를 창출하려는 입장에서 어떤 데이터가 있는지조차 알기 어렵다는 점은 큰 문제다.

원하는 데이터베이스를 찾았다고 해서 금방 사용할 수 있는 것도 아니다. 여러 데이터 중 띄어쓰기 하나만 잘못되어도 코드가 아예 작동하지 않을 수도 있기 때문이다. 도로명주소의 경우 도로명은 공백 없이 모두 붙여 쓰는 것이 원칙이다. 현재는 수정되었으나, 일례로 충남 아산시 '아산밸리남로 110번길'은 데이터베이스 상에 공백을 포함하고 있어 문제가 되었다. 공백이 없다는 가정하에 코드를 작성하였는데 이렇게 데이터베이스에 문제가 있다면 원하는 결과를 얻기 어려울 뿐만 아니라,

어떤 오류가 있는지 파악하고 이를 우회하기 위해 많은 노력을 기울여야 한다. 데이터베이스 수정을 요청해도 당장 수정이 이루어지는 경우는 드문 데다가, 정해져 있는 다음 데이터베이스 업데이트 주기까지 기다려야 하는 경우도 허다하다. 이렇게 데이터에 하나라도 문제가 있으면, 개발자는 데이터베이스를 신뢰할 수 없고, 나중에 있을 더 큰 오류를 방지하기 위해 데이터를 하나하나 검증해야 한다.

이름만 같고 실상은 다른 데이터도 있다. 시도별 생활업종 현황 통계에는 편의점 항목이 있고, 지자체 인허가 통계에는 휴게음식점(편의점) 항목이 있다. 쉽게 헷갈릴 수 있는 명칭이지만 휴게음식점은 편의점뿐만 아니라 슈퍼마켓, 생과일주스점, 카페 등 여러 항목을 포함하고 있어서 같은 항목으로 취급해서는 안 된다.

### 철저한 검증과 빠른 피드백 통해 데이터 편의성 높여야

역사의 큰 흐름에서 혁명 전후에는 큰 단절이 있는 것 같아 보이지만, 혁명의 한가운데에 있는 사람에게는 그렇지 않다. 혁명은 지지분하고 사소한 결정이 모여 나중에 규정되는 것이다. 때문에 우리는 당장 눈앞에 닥친 지지분한 데이터를 어떻게 잘 정제하고 더 쉽게 이용할 수 있도록 만들지 고민하고 실천해야 한다. 오류가 없는 데이터란 불가능하다. 앞으로 점점 더 많은 데이터를 수집하게 되면 개발자 개개인이 오류에 대응하기란 불가능해진다. 공공데이터 수집, 관리의 주체인 정부가 데이터를 더 철저히 검증하고 개발자의 요청에 빠르게 응답하여야 한다. 데이터를 구성하는 항목은 무엇인지, 분류 기준과 관련 법규는 무엇인지 이용자의 입장에서 친절하게 안내도 해야 한다. 유효숫자는 우리가 신뢰할 수 있는 숫자의 자릿수이다. 모호하고 추상적이기만 한 4차 산업혁명으로의 전환을 고민하기보다는, 3.0을 어떻게 하면 3.1로 바꿀지, 3.01을 어떻게 하면 3.02로 바꿀지 고민하는 노력이 더 큰 변화를 이끌어낼 것이다. 혁신의 유효숫자를 어디에 둘지 고민해보아야 할 시점이다. 