유엔의 빅데이터 품질검증 기준과 시사점: 빅데이터의 국가통계 활용을 중심으로

UN's Big Data Quality Criteria and Their Implications: Focusing on National Statistics

진재현 | 한국보건사회연구원 전문연구원
고금지 | 한국보건사회연구원 연 구 원

1. 들어가며

일반적으로 통계라고 하면 수치화된 통계 자료를 의미한다. 일부는 자료를 수집하고 분석하는 방법론을 다루는 과학으로서의 통계학이라는 학문체계로 이해하기도 하고, 표본으로부터 산출된 값을 뜻하는 통계량(statistic의 복수형)으로 이해하기도 한다. 그러나 우리가 흔히 통계라고 말하는 것은 처음에 언급한 수치화된 통계 자료, 즉 통계 수치를 의미한다.

통계법 제3조에 따르면 '통계란 통계작성기관이 정부 정책의 수립·평가 또는 경제·사회 현상의 연구·분석 등에 활용할 목적으로 산업·물가·인구·주택·문화·환경 등 특정의 집단이나 대상 등에 관하여 직접 또는 다른 기관이나 법인 또는 단체에 위임 위탁하여 작성하는 수량적 정보(통계작성기관이 내부적으로 사용할 목적으로 작성하는 경우는 제외)이다'라고 정의하고 있다.

위에서 언급한 특정한 집단은 다른 것과 구별될 수 있는 것들의 모임이다. 통계 수치가 의미를 가지려면 그 통계의 대상 범위인 집단이 명확하게 정의되어야 하며 특정한 집단에 대한 현상은 체계적인 숫자로 표현할 수 있어야 한다. 통계는 전체 집단 또는 부분 집단에 관한 사실을 객관적으로 나타내는 것이나 집단을 구성하는 개체의 개별적인 정보를 나타내는 것은 아니다. 즉 통계는 대상이 되는 집단의 특성을 파악하기 위한 수단이라고 할 수 있다. 따라서 통계는 알고자 하는 집단의 현재, 과거, 미래의 상태를 추측할 수 있게 하고 그 집단만의 고유한 특성을 파악하여 타 집단과 비교하는 것도 가능하게 한다.

한편, 최근 조사 환경이 빠른 속도로 악화되어 국가통계(national statistics) 작성은 심각한 위기에 놓인 실정이다. 1인 가구 및 맞벌이 가구

가 빠르게 증가하여 조사 대상에 대한 접촉 자체가 어려워지고, 개인 정보 노출 기피 등으로 인한 조사 거부는 더욱 심화될 것으로 예상된다. 이와 관련해 2013년 유엔유럽경제위원회(United Nations Economic Commission for Europe: UNECE)의 통계정보시스템 관리 전문가들은 HLG[1]와 함께 국가통계로서의 빅데이터 중요성에 대해 논의했고 국가통계 작성에 빅데이터를 활용하지 않으면 빅데이터로 생산된 통계들을 시기적절하게 사용하며 성장해 가는 민간 영역에 뒤처질 것이라고 경고하기도 하였다.

이와 같은 조사 환경의 변화에 따른 통계의 정확성 및 신뢰도 악화에 대한 우려와 함께 빅데이터와 같은 다양한 데이터를 활용해 국가통계를 작성해야 한다는 목소리도 높아지고 있다. 하지만 빅데이터는 경우에 따라 특정한 집단을 대표하기 어려운 측면이 있고 대용량이기 때문에 데이터 생산 단계에서 체계적으로 검증하기 어렵다는 점에서 품질에 대한 우려의 시선 또한 존재한다. 따라서 본고에서는 유엔(UN)의 빅데이터 품질 검증 권고안을 살펴보고 시사점을 고찰하고자 한다.

2. 빅데이터의 정의와 관련 국가통계 제도

빅데이터란 기존의 데이터베이스 관리 도구로는 데이터를 수집·저장·관리·분석할 수 없는 대량의 데이터 집합으로 정의되며[2] 데이터 형태에 따라 정형 데이터와 비정형 데이터, 생성 주체에 따라 기계·소셜·관계 데이터로 구분된다.

빅데이터는 소셜네트워크서비스(SNS)의 발달에 따른 비정형 데이터의 폭증, 클라우드 컴퓨팅을 통한 데이터 저장 및 처리 비용의 하락, 대용량·초고속 유무선 네트워크의 보편화, 그림자 정보(위치 정보, 검색 패턴, 접속 기록)의 증가, 사물 정보통신망 확산에 따른 저변 확대 등 다양한 배경적 요인으로 인해 등장했다. 이처럼 빅데이터 시대가 우리에게 빠르게 다가올 수 있었던 것은 원천 데이터의 다양성에 따라 새로운 가치를 창출하는 것이 가능해졌기 때문이다.

한편 국가통계란 정부와 관련된 기관에서 생산되어 정부뿐만 아니라 일반 이용자에게도 제공되는 통계로, 과거에는 유엔[3]에서 제시한 공식통계(official statistics)라는 용어를 많이 사용했으나 최근에는 통계청을 중심으로 국가통계라는 말을 주로 사용하고 있다.

1) HLG(High-Level Group for the Modernization of Statistical Production and Service): UNECE가 주관하는 유럽통계기관장회의(Conference of European Statisticians: CES) 산하 통계 생산과 서비스의 현대화를 위한 고위급 회의체.

2) Manyika, J. et.al.(2011), Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, p.1.

3) UNITED NATIONS STATISTICAL COMMISSION, et. al.(1994), Fundamental principles of official statistics. Official Records of the Economic and Social Council.

국가통계의 가장 주요한 기능은 사회 통합 기능과 정책 수립·평가 기능이다.[4] 우선 사회 통합 기능은 국가통계가 사회적 현상에 대해 신뢰할 수 있는 정보를 제공할 경우 사회구성원들은 이에 따라 공통적 인식을 형성하고 정부의 의사결정에 참여하도록 한다는 것이고, 정책 수립·평가 기능은 국가 선진화를 위한 효율적인 국정 운영을 가능하게 만드는 것을 의미한다.

우리나라는 대다수의 통계 생산에서 분산형 통계제도를 채택한 국가로, 각 통계작성기관(중앙행정기관, 지방자치단체, 지정 기관[5])이 업무 수행에 필요한 통계를 자체적으로 작성한다. 분산형 통계제도는 분야별 전문지식을 관련 통계 개발에 효율적으로 활용할 수 있으며 통계 수요에 신속하게 대응할 수 있다는 장점이 있는 반면 국가적인 관점에서 볼 때 통계의 중복 생산으로 인한 인력 및 예산의 낭비, 통계전문요원의 집중적 활용의 어려움 같은 문제가 있어 통계 품질의 저하를 야기하기도 한다.[6] 통계청은 국가통계 조정 기관으로서 이러한 문제를 최소화하기 위해 국가통계를 대상으로 품질 진단을 수행한다. 자체 품질 진단의 경우 통계청에서 지침서를 작성해 배포하면 각 통계작성기관에서 통계 작성 기획, 통계 설계, 자료 수집, 자료 입력 및 처리, 자료 분석 및 품질 평가, 문서화 및 자료 제공의 통계 작성 전반에 걸친 자가 진단을 직접 수행하여야 한다. 또한 정기 품질 진단을 통해 외부의 통계 전문가에게 각 국가통계의 품질 진단을 의뢰해 개선 과제를 도출하고 각 통계작성기관에서는 개선 과제를 이행함으로써 품질을 관리하고 있다.

현재 우리나라의 국가통계는 통계청장의 승인을 받아 각 통계작성기관에서 작성하도록 되어 있다.[7] 통계법 제18조는 '통계작성기관의 장은 새로운 통계를 작성하고자 하는 경우에는 그 명칭, 종류, 목적, 조사 대상, 조사 방법, 조사 사항의 성별 구분 등 대통령령으로 정하는 사항에 관하여 미리 통계청장의 승인을 받아야 한다'고 고시하고 있다. 2016년 10월 현재 국가승인통계는 397개 기관에서 979종이 작성되고 있으며 통계청이 이러한 국가승인통계를 조정, 심사, 품질 관리 중이다.

4) 김용환 외(2015), 국가통계 이해, 통계교육원.
5) 법률에 따라 설립된 법인 중 일정한 요건을 갖추어 통계청장이 지정하는 작성 지정 기관.
6) 한국행정연구원(2012), 국가정책지표체계 구축방안 연구.
7) 다만 통계청은 국가통계를 조사통계와 가공통계는 승인제, 보고통계는 신고제로 이원화하여 관리하기 위해 통계법을 개정 중이다.

표 1. 국가승인통계 현황(2016.10.5. 기준)

(단위: 개, 종)

기관 구분	작성 기관 수	작성 통계 수	통계 종류별		작성 형태별		
			지정	일반	조사	보고	가공
계	397	979	94	885	433	455	91
중앙행정기관	43	358	59	299	181	143	34
지방자치단체	260	448	17	431	149	261	38
지정기관	94	173	18	155	103	51	19

자료: 통계청, 국가통계포털 http://kosis.kr/serviceInfo/serviceinfo_0202List.jsp에서 2016.10.5. 인출.

3. 국가통계로 활용 가능한 빅데이터의 특징 및 유형

유엔이 제시한 국가통계 10대 원칙(fundamental principles) 중 제5조에 따르면 통계 작성 목적의 데이터는 다양한 유형의 원천 데이터를 활용할 수 있으며 통계작성기관은 품질, 적시성, 비용, 응답자 부담 등을 고려해 통계를 작성해야 한다고 하였다. 또한 UNECE는 (1) 의료 기록, 보험, 부동산, 조세 등의 행정 데이터 (2) 신용카드, 온라인 상거래(모바일 기기) 등의 상업용 및 거래 데이터 (3) 위성사진, 도로, 교통, 기후 등의 센서 데이터 (4) 모바일 휴대전화, GPS 등의 추적(tracking) 데이터 (5) 온라인 검색(상품, 서비스, 기타 정보)과 같은 행동 데이터 (6) 블로그, 소셜미디어 등의 의견(opinion) 데이터를 기존의 조사통계, 보고통계 형태로 작성된 국가통계를 개선할 데이터 원천으로 활용하도록 권고하였다.[8)]

그중에서도 행정 자료를 주된 원천 자료로 이

표 2. UNECE가 국가통계로 활용 가능한 것으로 제시한 빅데이터

구분	내용
행정 데이터	의료 기록, 보험, 부동산, 조세 등
상업용 및 거래 데이터	신용카드, 온라인 상거래(모바일 기기) 등
센서 데이터	위성사진, 도로, 교통, 기후 등
추적(tracking) 데이터	모바일 휴대전화, GPS 등
행동데이터	온라인 검색(상품, 서비스, 기타 정보)
의견(opinion) 데이터	블로그, 소셜미디어 등

자료: GLASSON, M., et. al.(2013), What does "Big Data" mean for Official Statistics, Paper for the High-Level Group for the Modernization of Statistical Production and Services. 내용 재정리.

8) GLASSON, M., et.al.(2013), What does "Big Data" mean for Official Statistics, Paper for the High-Level Group for the Modernization of Statistical Production and Services.

용해 작성하는 행정통계는 공공 영역에서의 활용도가 증가하고 있는 빅데이터의 한 유형이다. 행정 자료는 공공기관이 직무상 작성, 취득하여 관리하고 있는 문서, 대장 및 도면과 데이터베이스(DB) 등의 전산 자료로 정의(통계법 제3조 7호)하는데, 행정기관이 수집 및 처리 방법에 대한 권한을 가지고 있으며 주로 비통계적인 목적으로 모집단 전체를 포괄하여 수집된다는 특징이 있다. 조사 환경이 악화되고 조사 비용 및 응답 부담이 점점 커져 가는 상황에서 행정통계는 조사통계를 대체할 수 있는 좋은 대안이 된다. 한 예로 통계청의 임금근로일자리 행정통계는 건강보험공단, 국민연금공단, 고용노동부, 국세청, 대법원에 수집된 행정 자료를 성·연령별, 기업체 규모별, 산업 분류별로 분류해 작성한 통계로 정부 일자리 정책 수립과 취업 준비자의 일자리 선택에 필요한 기초 자료를 제공한다. 또한 통계청은 2015년 인구주택총조사에서 주민등록부, 건축물대장 등 인구, 가구, 주택과 관련된 행정 자료를 이용해 응답에 대한 국민의 부담을 줄이는 동시에 약 1455억 원의 예산도 절감했다. 하지만 이러한 행정통계는 행정 자료 취급 담당자의 통계 업무 숙련도 및 업무량에 따라 품질의 편차가 클 수밖에 없다는 특징이 있다.

또한 민간 영역에서는 SNS, 추적 데이터, 신용 정보 등의 빅데이터 활용 사례가 증가하고 있으며 그 효용성 역시 긍정적으로 평가된다. 행정 자료 이외의 빅데이터 활용 국가통계 작성의 한 예로 부산광역시 서비스인구이동통계를 들 수 있다. 부산시의 경우 일시적으로 도시를 방문하는 실제 인구가 증가함에도 불구하고 현재 사용하는 인구통계에서는 지속적으로 감소하는 것으로 나타나 도시 인프라 구축 등의 수요를 잘 반영하지 못한다는 지적에 대응하고자 SKT 자료를 활용해 부산광역시 서비스인구(상주 인구+비상주 인구)이동통계를 작성하였다. 부산광역시 서비스인구이동통계는 표본의 신뢰성 부족 등의 사유로 통계청으로부터 1회한으로 승인받은 바 있다. 상기 유형의 인구학적 통계는 빅데이터의 활용이 지금처럼 상용화되기 전에는 작성이 가능하지 않았으나 현재는 그 활용 가치가 크기 때문에 다양한 유형의 빅데이터를 국가통계 작성의 원천 데이터로 활용하는 시도의 한 예라 할 수 있다.

표 3. 빅데이터, 행정 자료 및 국가통계 특징 비교

특성	국가통계	빅데이터(행정 자료)	빅데이터(행정 자료 이외)
생산 이유	통계 생산 용도	행정 관리 목적	특정 비즈니스 목적
분석·처리 목적	모집단 추론 정책 효율성 지원	행정 관리 지원	비즈니스 목표에 부합 (모델링/최적화)
모집단 대표성	강함	강함(대상 집단)	약함
정보 추가	추가 가능	통계 생산 용도는 추가적인 보완 조치 필요	통계 생산 용도는 추가적인 보완 조치 필요
주 분석 기법	확률 기반 표본이론	빈도 분석 (빅데이터 기법 일부 활용)	데이터마이닝 기계학습 최적화
수집 비용 (자료 단위 기준)	높음	중간	일반적으로 낮음
수집 간격	주기적	거의 실시간	실시간

자료: 이지영(2015), 빅데이터의 국가통계 활용을 위한 기초연구, 통계개발원. 내용 재정리.

4. 유엔의 빅데이터 품질 검증 기준[9)]

UNECE는 앞에서 언급한 것처럼 국가통계 작성 기관에 국가통계 작성 시 빅데이터를 활용하라고 권고했으며 2014년 "통계 생산 현대화를 위한 빅데이터의 역할(The role of Big Data in the Modernization of Statistical Production)" 프로젝트를 통해 국가통계 활용을 위한 빅데이터 품질 기준을 제시하였다. 또한 국가통계로서 빅데이터의 적용 가능성을 파악하고자 호주, 스웨덴, 캐나다와 같은 통계 주요국에서 만든 품질 진단 체계를 사용해 품질 검증 3단계를 제시하였다. 품질 검증 3단계는 먼저 데이터를 수집하거나 수집 중인 상태를 의미하는 투입(input) 단계와 데이터의 가공, 분석 등이 수행되는 과정인 전환(throughput) 단계, 빅데이터에서 얻은 통계의 품질을 검증하는 산출(output) 단계로 구분된다. 그중 투입과 산출 단계는 출처(source), 메타데이터(metadata), 데이터(data)로 구성된 3가지 영역으로 구분되는데, 이것은 네덜란드 통계청에서 고안한 행정 데이터 품질 진단 체계의 개념에서 착안되었다.

가. 투입 단계

품질 검증 3단계 중 투입 단계는 데이터의 수

9) UNECE(2014), A Suggested Framework for the Quality of Big Data, Deliverables of the UNECE Big Data Quality Task Team.

집과 관련한 것으로 크게 출처, 메타데이터, 데이터 영역으로 나뉜다. 그중 수집 후 활용도가 생기게 되는 데이터의 경우 수집 즉시 사용이 가능하기도 하지만 데이터에 관한 정보만이 유용하게 사용될 수도 있다. 그런데 출처와 메타데이터는 데이터 수집 전에 데이터 관련 정보를 제공하는 특징이 있으므로 실질적으로는 데이터 수집 이전에 미리 품질 평가를 할 수 있다. 이 단계에서는 수집된 데이터의 적합 여부에 대한 검토와 품질 평가가 동시에 이루어지기도 한다.

투입 단계 중 출처 영역에서 다루는 품질 검증 요소에는 주로 데이터를 생산하는 기관의 영향력 파악 및 데이터 이용 가능성 확보 시 중요한 제도·사업적 환경, 그리고 데이터 사용 시 법적으로 요구되는 개인 정보 보호와 보안이 있다. 다음으로 메타데이터 영역은 데이터의 구조 및 형식 등을 다루는 복잡성, 메타데이터의 완전성, 데이터 사용을 위한 부가 자원의 유용성, 데이터의 시간 관련 요소, 다른 데이터 파일과의 연결 가능성, 변수 평가 및 변수 정의 등의 일관성, 타당성 등 데이터의 속성 정보를 검증한다. 마지막으로 데이터 영역에서는 데이터 그 자체의 정확성과 선별성, 연계 가능성과 일관성, 타당성 등을 검증한다.

표 4. 빅데이터 품질 검증 투입 단계 영역별 구조

영역	품질 검증 요소	고려 사항
출처	제도·사업적 환경 (Institutional/Business Environment)	지속적인 데이터 제공 가능 여부 /신뢰도/투명성 및 해석 가능성
	개인 정보 보호·보안 (Privacy and Security)	관련 법률/데이터 제공자(Data provider) 대 데이터 소유자(Data keeper) /이용 제한 사항
메타데이터	복잡성(Complexity)	데이터 처리에 대한 기술적 제한 요소/데이터의 구조화 여부 /데이터의 가독성/데이터의 계층 및 중첩 여부
	완전성(Completeness)	메타데이터의 활용 가능성, 해석 용이성, 완전성 여부
	유용성(Usability)	데이터 처리·저장·분석에 필요한 부가적인 자원/위험성 분석
	시간 관련 요소(Time-related factors)	시의성/주기성/시점별 개념 및 작성 방법의 변화
	연계 가능성(Linkability)	연계변수의 존재 여부와 품질/연계 수준(Liking level)
	일관성(Coherence-consistency)	표준화된 주요 변수 개념 /주요 변수에 대한 메타데이터의 이용 가능 여부
	타당성(Validity)	통계 산출 방법론과 과정의 투명성 및 공정성
데이터	정확성(Accuracy and selectivity)	총표본오차/참고 데이터 세트/선택성
	연계 가능성(Linkability)	연계변수의 질
	일관성(Coherence-consistency)	메타데이터와 관측된 데이터값의 일관성
	타당성(Validity)	통계 산출 과정·방법과 관측된 데이터값의 일관성

자료: UNECE(2014), A Suggested Framework for the Quality of Big Data, Deliverables of the UNECE Big Data Quality Task Team. 내용 재정리.

나. 전환 단계

다음으로 수집된 데이터를 가공하고 분석하는 전환 단계는 데이터 품질에 대해 시스템의 독립성(system independence), 데이터 세트의 안정 상태(steady states), 품질 게이트(quality gate) 등 총 3가지 기본 원칙을 제시하고 있다. 먼저 시스템의 독립성 원칙은 데이터를 분석하고 변형하는 것이 특정한 프로그램에 독립적이어야 한다는 것이다. 다음으로 데이터 세트의 안정 상태는 데이터 생산과 분석에 종사하는 모든 사람이 데이터 품질을 객관적이고 명확하게 이해하는 상태를 의미하는 것으로, 복잡한 데이터 생산 과정에서 필수적이라고 할 수 있다. 마지막으로 품질 게이트는 데이터 분석 과정에서 데이터 품질을 명확히 평가하기 위해 엄격한 품질 보증 과정을 두는 것을 의미한다. 품질을 평가하기 위해 사용되는 측정 도구와 품질 평가가 이루어지는 품질 게이트의 지점은 반드시 사전에 정해져야 한다.

다. 산출 단계

산출 단계의 품질 검증 체계에서는 데이터 소비자에게 전달되는 최종 데이터의 품질을 평가하며 데이터 보급의 적절성과 데이터 투명성에 초점을 둔다. 본 단계에서의 품질 검증은 투입 및 전환 단계에 비해 전반적이고 포괄적인 검증 경향을 띠며 앞서 언급한 투입 단계에 해당하는 품질 검증 체계와 유사하여 투입 단계 품질 검증 체계들을 재검토한다는 특징이 있다.

표 5. 빅데이터 품질 검증 산출 단계 영역별 구조

요소	품질 검증 요소	고려 사항
출처	제도·사업적 환경 (Institutional/Business Environment)	빅데이터 출처의 종류 데이터 전송을 위한 준비 및 품질 보증
	개인 정보 보호·보안 (Privacy and Security)	관련 법률 데이터 사용 시 실질적 문제점
메타데이터	복잡성(Complexity)	데이터 처리 과정/결과물의 제한 사항
	접근 가능성 및 명확성 (Accessibility and Clarity)	데이터 및 메타데이터의 가용 여부 정의 및 설명의 명확성 표준의 적격 여부
	연관성(Relevance)	데이터 사용 목적과의 관련성
데이터	정확성(Accuracy and selectivity)	전통적 정확성 측정법(예: 표준오차, 편향 등)/선택성
	타당성(Validity)	유사 지표와의 상관관계 /유용성/개념적 건전성
	시간 관련 요소(Time-related factors)	시의성/주기성

자료: UNECE(2014), A Suggested Framework for the Quality of Big Data, Deliverables of the UNECE Big Data Quality Task Team. 내용 재정리.

5. 국내 빅데이터 관련 주요 정책 동향

가. 빅데이터 활용 및 공공데이터 관련 정책 동향

정부는 미래창조과학부의 인공지능 및 기계학습을 적용한 빅데이터 선도 사례 육성 사업과 행정자치부의 공공데이터 정책을 통해 빅데이터 정책을 주도하고 있다. 미래부는 공공과 민간 영역의 융합을 위한 사업을 주로 추진하는데, 최근에도 지난 5월에 한국정보화진흥원(NIA)과 함께 빅데이터의 확산을 위한 2016년 선도 시범 부문의 4개 과제와 산업 확산 부문의 2개 과제를 선정한 바 있다.

그 내용을 살펴보면 안전 분야에서는 통신 로밍 빅데이터를 활용해 메르스, 지카바이러스 등과 같은 해외 감염병의 국내 유입 차단을 추진하고, 유통 분야에서는 티-커머스(T-commerce) 편성에 딥러닝을 이용한 마케팅을 통해 매출 향상에 기여하며, 제조 분야에서는 자동차 전자부품 생산 시 4M(Man, Machine, Material, Method) 데이터와 작업 환경 정보, 검사 영상 정보를 융합 분석하여 품질 향상을 추진하는 등 미래부는 민간 영역의 빅데이터를 활용해 새로운 경제적 가치를 창출할 수 있는 사업을 중점적으로 추진 중이다.

표 6. 미래부-NIA 선정 2016년 빅데이터 시범 사업

구분	주관 기관/참여 기관	과제명
선도 시범	KT/질병관리본부	로밍 빅데이터를 활용한 해외 유입 감염병 차단 서비스
	더블유쇼핑/한동대	빅데이터 딥러닝 기술 활용 스마트 티-커머스 서비스 개발
	매일유업/한국그린비즈니스협회	유가공 업종 제조 생산·에너지 최적화를 위한 빅데이터 플랫폼 개발
	유라/충북대학교 등	딥러닝 기술 기반의 대용량 제조 데이터 분석 서비스 플랫폼 개발
산업 확산	ING생명/생명보험협회	생명보험 빅데이터 전략 모델 개발 및 확산
	삼성중공업/현대중공업 등	제조업 빅데이터 전략 모델 개발 및 실증

자료: 미래부·한국정보화진흥원 보도자료(2016. 5. 31.), 미래부-NIA, 2016년 빅데이터 시범 사업 착수.

한편 정부의 빅데이터 관련 공공데이터 정책은 행자부와 공공데이터전략위원회(Open Data Strategy Council)를 중심으로 추진되고 있다. 공공데이터전략위원회에서는 공공데이터에 관한 정부의 주요 정책과 계획을 심의, 조정하고 추진 사항을 점검하고 평가하며, 행자부에서는 전

략위원회에 상정하기 위한 제1차 공공데이터 기본계획(2013.~2016.6.)에 대한 평가와 함께 제2차 공공데이터 기본계획(2017.~2019.)을 수립해 공공데이터 정책을 주관 중이다.

해외의 공공데이터 정책 기조가 '정부 투명성(transparency)'에서 '경제적 효과 창출(economic benefit)' 수단으로 변화하는 가운데, 유럽연합(EU)은 2009년부터 EU 국가들의 데이터 개방을 촉진하였고 2015년부터는 EU 국가들에 데이터를 활용한 경제 효과 창출을 독려하고 있다. 이에 발맞추어 행자부에서는 공공데이터를 경제활동 수단으로 활용하고, 민간과의 협업을 확대하기 위한 제2차 공공데이터 기본계획을 수립 중이다.

표 7. 제2차 공공데이터 기본계획 수립을 위한 5대 패러다임 대전환 전략

정책 주도		제1차 기본계획	제2차 기본계획(안)
		• 정부 주도	▶ 국민/기업과 협력
패러다임 대전환	개방	• 양적 개방	▶ 질적 개방
	기반	• 수직적 플랫폼	▶ 수평적 공유(네트워크) 플랫폼
	생태계	• 창업 중심 생태계	▶ 융합산업 중심 생태계
	역량	• 제도 정비 및 홍보 중심	▶ 신기술 및 상품화 중심
	효과	• (개방/활용 수)양적 중심	▶ 사회·경제적 파급효과 중심

자료: 행정자치부(2016), 「제2차 공공데이터 제공 및 이용활성화」 기관별 기본계획 작성 지침.

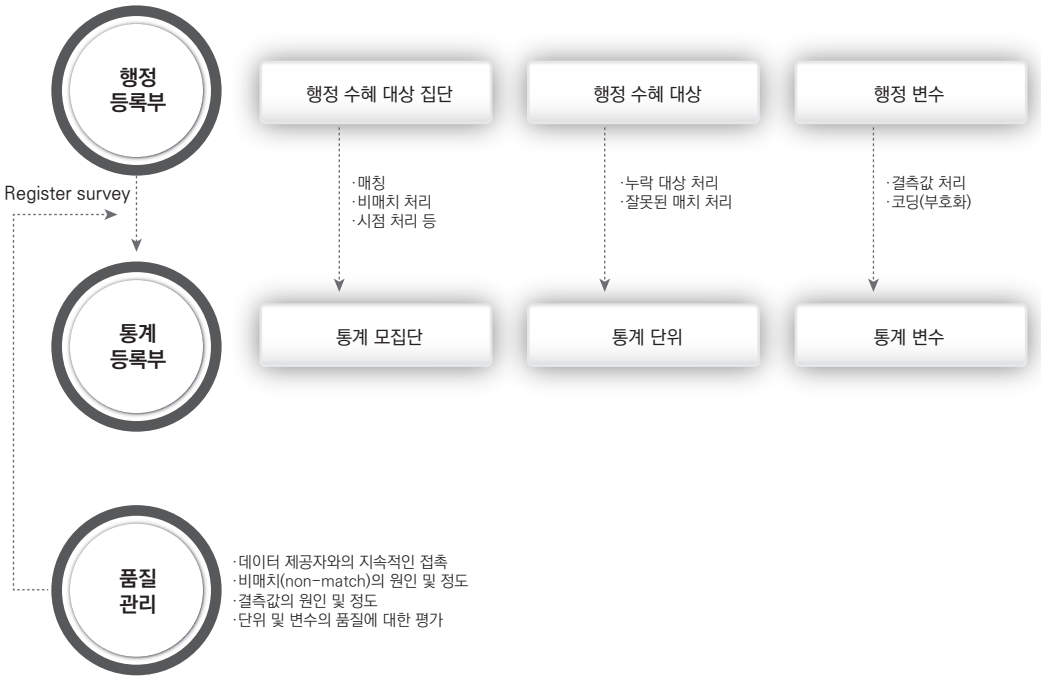
정부3.0의 기조하에서 미래부와 행자부를 중심으로 한 정부의 빅데이터 정책은 빅데이터를 활용해 새로운 비즈니스 기회를 창출하고 다양한 서비스를 개발해 창조경제의 토대가 되도록 하는 데 지향점이 있다.

나. 빅데이터 관련 통계제도 동향

국가통계기관인 통계청에서는 각 행정기관에서 보유한 행정 자료의 활용도를 높이기 위해 인구, 사업체·기업체, 고용, 주택 등 다양한 행정 자료를 각 기관으로부터 확보해 개별 통계등록부를 구축하고 이를 연계한 종합 통계등록부를 구축하는 업무를 수행하고 있다. 종합 통계등록부는 인구 DB(성별, 연령, 가족관계 등), 사업체·기업체 DB(사업체 생성 및 소멸 등), 부동산 DB(건물 및 주거지 정보 등), 고용 DB(개인 소득, 고용 형태 등)를 주소, 개인식별번호, 사업자등록번호로 연계하여 각 DB 개별적으로는 확보

할 수 없는 새로운 정보를 제공하고 있다. 통계청에서는 통계등록부를 구축하면서 비매치, 결측값 등 여러 데이터를 연계해 발생할 수 있는 품질 저하 요소에 대해서는 별도의 관리 및 평가를 하지만, 각 기관으로부터 수집된 행정통계의 품질을 평가하기는 상대적으로 쉽지 않은 상황이다.

그림 1. 통계청 통계등록부 시스템 개요



자료: 은희훈(2016), 행정자료를 활용한 통계생산방법, 제6회 국가통계방법론 심포지엄 발표자료, 재정리.

한편 통계청에서는 현재 빅데이터를 활용하면서 표본 대표성, 오차 산출 불가능 등의 문제로 인해 국가통계로 승인하기 곤란하나 정책적으로 작성·공표할 필요가 있을 경우에는 '1회한 통계승인 제도'를 통해 국가통계로 관리하고 있다. 그러나 제11차 국가통계위원회(2015. 10. 29.)에 상정된 통계관리체제 개편(안)에 따르면 1회한 통계승인제도는 기준이 불명확하기 때문에 폐지하고, 통계작성기관에서 자체 관리하는 방향으로 통계법 개정을 추진 중이다.

6. 나가며

국가통계의 품질은 국가 정책을 수립하는 데 다방면으로 중대한 영향력을 끼친다. 통계 생산 과정에서의 오류는 곧바로 정책 수립의 오류로 연결되므로, 통계작성기관은 통계 생산 과정의 엄밀성을 위해 최선의 노력을 다해야 한다. 각 통계작성기관은 유엔에서 권고했고 앞에서 언급한 여러 정부 및 민간기관의 빅데이터 활용 사례들처럼 변화하는 국가통계 작성 환경에 대한 대응으로 빅데이터와 같은 다양한 유형의 데이터를 활용해 국가통계를 작성하기 위해 노력하여야 한다.

빅데이터 관련 산업은 빠르게 변화하고 있기에 그 개념 및 범위의 설정조차도 어려운 것이 현실이지만 품질 검증이 선행되지 않으면 그것은 대량의 폐기물(Big Garbage)이 될 수밖에 없다. 더욱이 빅데이터는 통계 작성을 목적으로 수집된 것이 아니기 때문에 국가통계로 활용하기 위해서는 담당자가 아니면 그 작성 과정을 알 수 없는 빅데이터에 대한 품질 검증이 반드시 선행되어야 할 것이다.

앞에서 유엔이 제시한 빅데이터 품질 검증 기준을 통해 데이터를 수집하거나 수집 중인 상태를 의미하는 투입 단계, 데이터의 가공, 분석 등이 수행되는 과정인 전환 단계, 빅데이터에서 얻은 통계의 품질을 검증하는 산출 단계에서 고려해야 할, 통계 생산 과정 전반에 대한 품질 검증 체계를 살펴보았다. 빅데이터 품질 검증을 통한 국가통계 활용도 제고를 위해 데이터 수집, 정제, 분석, 전파 등의 통계 작성 과정 전반을 검증할 수 있는 국내 빅데이터 보유 실정을 반영한 품질 검증 기준에 대한 연구가 필요하다. 그리고 빅데이터를 보유하고 있는 각 기관은 대표성, 신뢰성 등의 한계를 지닌 빅데이터 통계를 공개 또는 활용할 때 통계 전문 연구기관에 컨설팅 및 정보 제공에 대한 자문을 적극적으로 할 필요가 있다.