빅데이터 품질관리 표준화 현황

이창수 _ 강릉원주대학교 산업경영공학과 교수
김선호 _ 명지대학교 산업경영공학과 교수
이진우 _ 투이컨설팅 부사장

1. 머리말

2020년까지 1인당 1초에 1.7MB씩 데이터를 만들어내고 있다고 한다. 1인당 일주일에 1TB의 하드디스크가 필요할 정도로 데이터가 생성된다는 얘기다. 일 년에 50TB가 넘는 데이터가 생긴다니 이쯤 되면 데이터의 양이 기하급수적으로 늘고 있다는 표현이 조금이나마 실감 난다[1].

기업의 입장에서 이렇게 많아지는 데이터를 저장해서 처리하는 일은 보통 일이 아니다. 특히 신규 시스템 도입 등의 이유로 기존 데이터를 이전(migration)해야 하는 상황이 발생하는데, 기존에 사용하고 있는 시스템 개수가 많은 경우 어려움이 가중된다. 그러나 이보다 더 어려운 문제는 데이터 품질이라는 조사 결과가 나왔다. 데이터 품질 문제는 프로젝트 지연, 데이터 중복, 데이터 품질 저하, 비용 증가 등의 결과를 초래한다[2].

가트너는 데이터 품질 저하로 비즈니스 가치가 떨어진다고 지적하면서 매년 평균 1,500만 불의 손실이 발생한다고 추정하였다. 이러한 재정적 손실 이외에 고객의 불신, 경쟁력 상실, 디지털 주도권 약화를 초래한다고 언급하고 있다. 반면 에어비앤비와 아마존 같은 혁신 기업들은 우수한 데이터 품질을 통해 고객이 어떠한지, 어디 있는지, 무엇을 좋아하는지를 파악하고 있다[3].

이와 같이 데이터의 양은 엄청난 속도로 증가하고 있으며 늘어나는 데이터를 처리하는 데 직면한 가장 큰 어려움은 데이터의 품질을 확보하는 것이다. 성공하는 혁신 기업은 데이터 품질 문제를 적절하게 다루고 있다는 것을 알 수 있다.

데이터의 품질을 다룬 최초의 연구는 1972년 스웨덴 왕립공대 Kristo Ivanov의 박사학위 논문이다. 그의 '정보 품질 관리: 데이터 은행 및 경영 정보 시스템에서 정보의 정확성 개념'이라는 제목의 논문에서는 완전성, 보안성, 신뢰성, 유효성, 정확성 등 데이터 품질 특성을 다루고 있다[4].

데이터 품질관리에 대한 국제 표준화는 ISO TC 184/SC 4(산업 데이터) 산하에 데이터 품질 연구반인 WG 13이 2006년에 결성되면서 개발이 시작되었다. WG 13은 ISO 8000(데이터 품질) 프로젝트를 통해 마스터 데이터, 제품 데이터, 일반 데이터를 대상으로 표준을 개발하고 있다. 미국, 스웨덴, 독일, 노

르웨이, 프랑스, 영국, 한국이 주로 개발에 참여하고 있다. 특히 한국은 ISO 8000-150, ISO 8000-61 등 핵심 표준을 포함하여 현재 개발 완료되었거나 개발 중인 22개 표준 중에서 7개 표준을 개발함으로써 데이터 품질 표준의 선도적 역할을 담당하고 있다. 이러한 데이터 품질 표준 개발 능력을 기반으로 스페인과 국제 협력을 통해 IoT 데이터 품질 표준을 개발하고 있으며 제조 데이터 품질 표준을 개발하는 등 산업 도메인을 지속적으로 확대해 나가고 있다.

2. 표준화 추진 동향

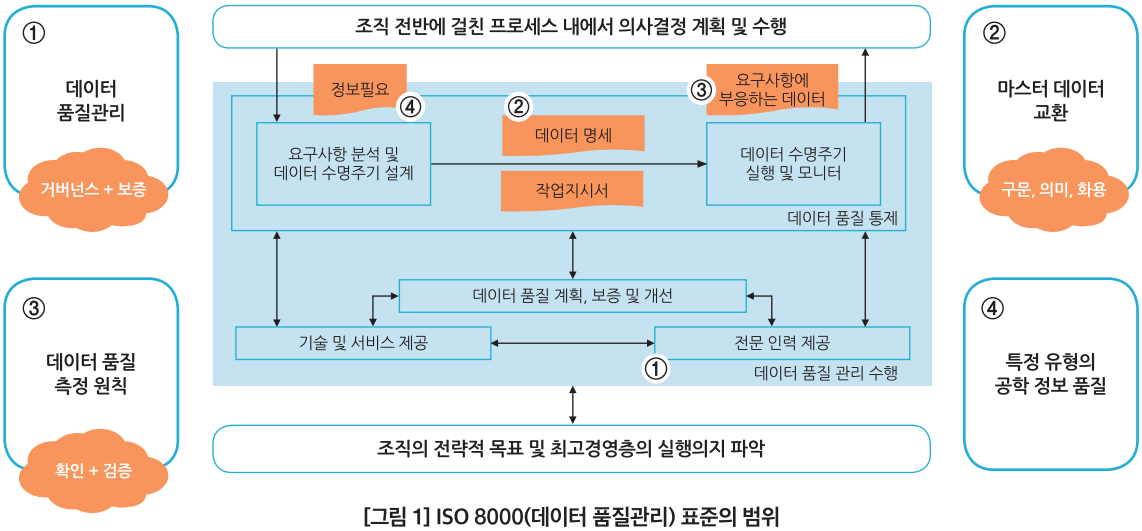
2.1 ISO 8000 개발 현황

데이터 품질 국제 표준은 ISO TC 184/SC 4/WG 13에서 진행 중인 ISO 8000(데이터 품질) 프로젝트로 개발된 표준이 대표적이며 2019년 2월 현재 15개의 데이터 품질 표준이 제정 완료되었고 7개 표준이 개발중에 있다. ISO 8000은 초기에는 ISO 22745에 의해 운영되는 기술사전 및 카탈로그 데이터의 품질

<표 1> ISO 8000 데이터 품질 표준 개발 현황

[2019년 2월 현재]

구분	Part	제목	설명	발행년도
개념	1	Overview	개요	2011
	2	Vocabulary	용어	2018
	8	Information and data quality: Concepts and measuring	정보 및 데이터 품질: 개념 및 측정	2015
일반 데이터	51	Data governance: Exchange of data policy statements	데이터 거버넌스: 데이터 교환 정책 선언문	개발중
	60	Data quality management: Overview	데이터 품질관리: 개요	2017
	61	Data quality management: Process reference model	데이터 품질관리: 프로세스 참조 모델	2016
	62	Data quality management: Organizational process maturity assessment: Application of standards relating to process assessment	데이터 품질관리: 조직 프로세스 성숙도 평가: 프로세스 평가와 관련된 응용 표준	2018
	63	Data quality management: Process measurement	데이터 품질관리: 프로세스 측정	개발중
	64	Data quality management: Organizational process maturity assessment: Application of the Test Process Improvement method	데이터 품질관리: 조직 프로세스 성숙도 평가: 테스트 프로세스 개선 방법의 응용	개발중
	65	Data quality management: Process measurement questionnaire	데이터 품질관리: 프로세스 측정 질문지	개발중
	66	Dataqualitymanagement:Assessment indicators for data processing in manufacturing operations	데이터 품질관리: 제조 운영에서 데이터 처리를 위한 평가 지표	개발중
	81	Data quality assessment methods: Profiling	데이터 품질 평가 방법: 프로파일링	개발중
마스터 데이터	100	Master data: Exchange of characteristic data: Overview	마스터 데이터: 특성 데이터 교환 : 개요	2016
	110	Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification	마스터 데이터: 특성 데이터 교환 : 구문, 의미 코딩, 데이터 명세 적합성	2009
	115	Master data: Exchange of quality identifiers: Syntactic, semantic and resolution requirements	마스터 데이터의 품질 식별자의 교환: 구문, 의미 및 변환 요구사항	2018
	116	Master data: Exchange of quality identifiers: Application of ISO 8000-115 to authoritative legal entity identifiers	마스터 데이터: 품질 식별자 교환: ALEI에 대한 ISO 8000-115 응용	개발중
	120	Master data: Exchange of characteristic data: Provenance	마스터 데이터: 특성 데이터 교환 : 출처	2016
	130	Master data: Exchange of characteristic data: Accuracy	마스터 데이터: 특성 데이터 교환 : 정확성	2016
	140	Master data: Exchange of characteristic data: Completeness	마스터 데이터: 특성 데이터 교환 : 완전성	2016
	150	Master data: Quality management framework	마스터 데이터의 품질관리 프레임워크	2011
제품 데이터	311	Guidance for the application of product data quality for shape (PDQ-S)	ISO 10303-59(PDQ-S) 응용 지침	2012



[그림 1] ISO 8000(데이터 품질관리) 표준의 범위

에 초점을 두고 개발되기 시작하였으나 범위가 점차 일반 데이터, 마스터 데이터, 제품 데이터로 확대되었다. ISO 22745는 나토 코드 체계에 기반을 둔 마스터 데이터 정의 및 교환 표준이다. 마스터 데이터 분야인 ISO 8000-100 시리즈와 제품 데이터 분야인 ISO 8000-311은 데이터 중심(data-centric)의 품질 개선 방법을, ISO 8000-60시리즈와 ISO 8000-150은 프로세스 중심(process-centric)의 품질개선 방법을 제시하고 있다. WG 13에서 ISO 8000 표준으로 개발되는 현황은 〈표 1〉과 같으며 표준의 범위는 [그림 1]과 같다.

조직에서 내리는 의사결정에 필요한 정보가 발생하면(④) 조직이 원하는 조건이 어떤 것인지를 살펴보고(②) 해당 데이터의 형식을 갖추어 조직에 제공한다(③). 이 과정에서 데이터의 품질 특성을 일정 수준 이상 갖추고 있는 데이터를 확보하고(④) 쌍방간에 원하는 의미를 명확하게 전달하여(②) 해당 시점에서 조직이 필요로 하는 데이터를 제공한다(③). 데이터 품질관리는 데이터 품질계획을 세우고 보증하는 체계를 제공한다(①).

2.2 기타 데이터 품질 표준

데이터 품질은 도메인별로 개발되기도 하는데, 소프트웨어의 데이터 품질, 지리정보시스템의 데이터 품질, 교통정보시스템의 데이터 품질 표준이 개발되었다.

ISO/IEC JTC 1/SC 7(소프트웨어 및 시스템 공학)은 소프트웨어 제품 품질 요구사항 및 측정 표준의 일환으로 데이터 품질 모델 ISO/IEC 25012 표준을 제정하여 정확성, 완전성 등 15개의 데이터 품질 특성을 정의하고 있으며, ISO/IEC 25024는 ISO/IEC 25012에서 정의한 데이터 품질을 정량적으로 측정하는 방법을 보여준다.

지리 정보 표준을 개발하는 ISO/TC 211은 지리정보시스템의 데이터 품질인 ISO 19157을 개발하여 데이터 품질 요소, 데이터 품질 측정, 지리 데이터 품질 평가 절차, 데이터 품질 보고 원칙 등을 정의하고 있다.

지능 교통 정보 표준을 개발하는 ISO/TC 204는 ISO 21707 표준을 개발하여 지능 교통 정보시스템에서 데이터 공급자와 소비자 사이에 교환되는 데이터의 품질을 정의한다. 이 표준은 교통 및 여행 정보

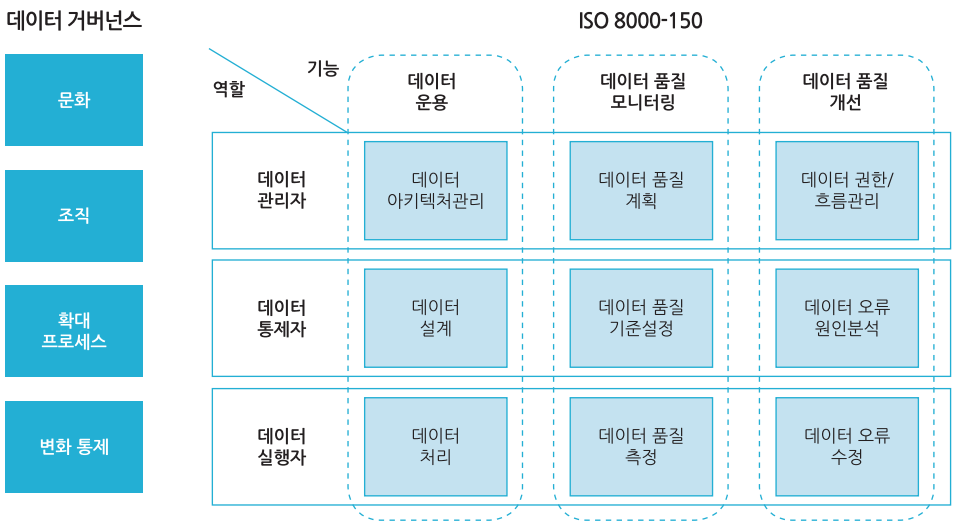
서비스와 교통 관리 및 통제 시스템에 적용되며 정확성, 시의성 등의 데이터 품질 특성을 정의한다.

데이터 품질 관련 단체로는 CMMI Institute, DAMA, EDM Council 등이 있다. CMMI Institute는 소프트웨어 품질 보증 기준으로 개발한 능력 성숙도 통합 모델(Capability Maturity Model Integration)을 기반으로 데이터의 품질 보증 기준인 DMM(Data Management Maturity)을 개발하였으며, 여기서 파생되어 나온 EDM Council은 성숙도 개념을 배제하고 능력도 중심의 모델인 Data Management Capability Assessment Model (DCAM)을 제시하고 있다. DAMA International은 데이터 관리 종합 지침인 DAMA-DMBOK(Data management body of Knowledge)를 발행하였다. 한국데이터산업진흥원은 데이터 품질관리 지침, 데이터 품질관리 성숙모형 등을 개발하여 국내 데이터 품질관리 표준을 선도하였으며 이를 토대로 ISO 8000 표준화 활동이 전개되었다.

2.3 데이터 품질 표준 적용 사례

ISO 8000-61은 데이터 품질관리에 필요한 프로세스 참조 모델을 제공한다. 20개의 프로세스를 Plan(데이터 품질 계획)-Do(데이터 품질 통제)-Check(데이터 품질 보증)-Act(데이터 품질 개선)의 선순환 구조를 갖도록 구성하여 프로세스의 능력도를 평가하고 개선하거나 데이터 품질관리에 대하여 조직의 성숙도를 개선하는 참조로 사용된다. 특히 ISO 8000-61은 프로세스 기반의 데이터 품질 개선을 제시함으로써 조직 차원에서 업무 프로세스와 연동하여 데이터 품질을 개선할 수 있다. 한국정보화진흥원은 ISO 8000-61을 근간으로 하여 공공데이터 품질관리 수준평가 사업을 수행 중에 있으며 공공데이터 품질관리 매뉴얼을 개발하여 보급하고 있다.

ISO 8000-150은 마스터 데이터에 대한 데이터 관리자의 역할과 책임에 대한 3x3 매트릭스 품질 관리 프레임워크를 정의한다. 영국의 철도 인프라를 운영하는 Network Rail은 인프라 자산에 대한 데이터를 물리적 자산과 동일하게 취급하여 확보, 유지보



[그림 2] ISO 8000-150을 데이터 거버넌스에 적용한 사례

수, 갱신, 처리하는 데 ISO 8000-150을 적용하여 관리하고 있다. 이는 데이터를 자산으로 취급하는 전략적 목표를 지원하는 것이다. 또한 영국의 DPA는 ISO 8000-150을 데이터 거버넌스로 확장하여 [그림 2]와 같은 개념을 제시하였다.

미국 전자상거래코드관리협회(ECCMA)는 전자카탈로그 구축 기술을 바탕으로 ISO 8000-110 표준을 보급하여 마스터 데이터 품질 관리자 인증 사업을 진행하고 있다. ECCMA는 데이터 품질 프로젝트를 개방형 기술사전 표준(ISO 22745, ISO 29002)으로 확장하였으며, 상업적으로 적용하는 데 많은 기여를 하고 있다.

국내 기업인 지티원㈜은 ISO 8000-61을 기반으로 연결기능이 강화된 스마트 제품(SCP)과 IoT 환경에서 자동 데이터 프로파일링, 데이터 감사, 데이터 규칙 관리, 데이터 품질 분석, 데이터 품질 분석 결과 보고 등을 제공하는 IoT 실시간 데이터 품질관리 솔루션을 개발하고 있다. 또한 ㈜위세아이텍은 ISO 8000-81을 기반으로 인공지능 기반의 빅데이터 분석과 머신러닝 프로세스, 데이터 품질관리 및 빅데이터 분석 플랫폼을 개발하고 있다.

2.4 국내 추진 체계

ISO 8000 프로젝트 초기부터 개발에 참여한 명지대학교 김선호 교수, 투이컨설팅 이진우 부사장, 한국데이터산업진흥원 임성준 팀장이 데이터 품질 표준을 선도하고 있으며 국가기술표준원 산업데이터전문위원회, ㈔스텝센터, 한국정보통신기술협회 PG606(메타데이터 프로젝트그룹)을 통해 데이터 품질 표준 개발이 이루어지고 있다.

한국정보화진흥원은 공공데이터포털을 통해 공공데이터 품질관리가 국제표준에 적합하도록 지원하고 있으며 한국데이터산업진흥원은 데이터 품질인증제도를 운영하면서 국제 규격을 선도하는 데이터 품질 방법론을 개발하고 있다.

3. 빅데이터 품질관리 표준화의 향후 이슈

■ 스마트 제조, IoT, 인공지능

ANSI/ISA-95에 기반을 두고 있는 IEC 62264는 독일이 주도하고 있는 인더스트리 4.0 모델의 중요한 축을 이루고 있다. IEC 62264는 제조 운영 관리(MOM, manufacturing operations management) 표준으로 생산활동 시 필요한 데이터를 정의하고 있다. 생산 과정에 수반되는 데이터의 품질을 데이터 처리 관점에서 보는 것은 ISO 8000-66이다.

■ 데이터는 돈이다

데이터 품질 저하로 인한 비용 발생은 수동적 의미이며 데이터로 수익을 창출하는 것은 적극적 의미이다. 데이터 경제학은 데이터에 경제적 중요성을 확고히 하여 업무에 수익구조를 만들고, 데이터를 실제 자산으로 관리하며 측정하는 프레임워크를 제공한다. 조직의 데이터 자산 가치와 정규 데이터 자산관리 프랙티스를 정량화하는 기초와 방법을 제공한다. IT 투자에 대한 수익률(ROI, return on investment)이 IT 투자에 대한 정당성을 확보해 주는 근거가 된다면 마찬가지로 데이터 품질에 대한 ROI를 적극적으로 고려해야 하는 상황이다. 즉, 데이터로 수익구조화를 고민할 때이다.

■ CDO의 등장

2018년 6월 미국 버지니아 주지사가 주정부 내에 최고 데이터 책임자(CDO, chief data officer)를 신설하는 법안에 서명하였다. 이는 정부 데이터 수집 및 보급법의 일부로 CDO의 역할은 데이터 사용, 저

장, 개인정보보호에 관한 지침 개발, 주정부 차원의 데이터 공유 활성화, 대국민 서비스 개선에 데이터 활용을 촉진 시키는 것이다. 이로써 미국은 총 18개 주에 CDO 역할을 수행하는 자리가 생기게 되었다. 구인구직 사이트인 Indeed.com에는 2019년 2월 현재 CDO에 대한 구인 광고가 1만 개가 넘게 올라와 있다.

■ 개인정보보호

새로운 개인정보보호규정(GDPR, General Data Protection Regulation)이 지난 2018년 5월부터 유럽연합에서 발효되었다. 이 규정을 심각하게 어길 경우 전 세계 연간 매출액의 4% 또는 2,000만 유로 중 더 높은 금액을 과징금으로 내야 한다. 포브스는 데이터 품질 없이 이 규정을 준수하는 것이 불가능하다고 말하고 있다[5].

4. 맺음말

데이터의 양이 늘어날수록 신뢰할 만한 데이터에 대한 필요성이 커지고 있는 것이 현실이다. 정형이든 비정형이든 믿을 만한 데이터에 대한 갈증이 높아지고 있다. ISO 8000 표준은 ISO 9000 표준의 틀인 선순환 구조를 원용하였다. 계획한 것을 실행하고 실행한 것을 점검하고 점검 결과를 개선에 반영하여 계획을 수정하는 사이클 구조이다. 또한 데이터값 자체에 대한 품질뿐만 아니라 데이터 품질 업무를 기업의 전체 업무 프로세스에 포함되도록 프로세스 기반으로 데이터 품질을 관리할 수 있는 방법을 제시하였다.

데이터 품질 표준은 IoT, 제조업, 지리정보, 소프트웨어 등의 분야로 지속적으로 적용을 확대하고 있다. 그만큼 각 분야에서 늘어나는 데이터의 양에 따라 데이터 품질의 중요성이 필수적이 되고 있다는 뜻이기도 하다. 정부에서도 공공데이터에 데이터 품질 표준을 적용하여 대국민 신뢰를 높이고 있으며 민간 부문에서도 데이터품질 인증제도가 확산되고 있다.

데이터를 보유한 조직은 데이터 처리에 데이터 품질 프로세스를 업무에 포함시켜 근본적인 품질관리가 될 수 있도록 구조적인 데이터 품질관리 방안을 도입할 필요가 있다. 한국의 데이터 품질관리 표준 개발 역량은 ISO 내에서도 우월한 위치에 있으며 후속 표준이 지속적으로 개발중에 있다. 이러한 표준 개발 성과가 각 산업 영역에 보급되어 적용할 수 있는 협업이 필요하다.

※ 본 연구는 2018년도 산업통상자원부 국가기술표준원 및 한국산업기술평가관리원의 지원을 받아 수행된 연구임[No. 10058970, 스마트제조 지원을 위한 제품 데이터 및 품질 정량화 표준 개발과 국제 표준 등록].

[참고문헌]

[1] Data Never Sleeps 6.0, Domo, 2018.
[2] The 2018 global data management benchmark report, Experian, 2018.
[3] Gartner's Data Quality Market Survey, Gartner, 2017.
[4] Ivanov, Kristo (1972). Quality-control of information: On the concept of accuracy of information in data-banks and in management information systems. Stockholm: The Royal Institute of Technology KTH(Doctoral dissertation, 258 pages).
[5] 4 Steps To A Data Quality Approach For Complying With New Data Regulations, 2018.5.22., Forbes.
[6] 김선호, 이창수, 정승호, 김학철, '공공데이터 품질관리를 위한 조직 성숙도 평가 모델', 정보화정책, 2015.
[7] 김선호, 이진우, 이창수, '활동능력수준 기반의 공공데이터 품질관리 성숙수준 평가 모델', 정보화정책, 2018.
[8] 임성준, 이정현, 이창한, 김정연, '데이터 프로파일링 기반의 데이터 품질평가 방안', 대한산업공학회 추계학술대회, 2018.