DQ인증 소개 및 현황

허준호 TTA AI데이터품질팀 수석연구원

1. 머리말

데이터가 디지털 혁신의 핵심 자산으로 자리 잡으면서, 그 품질의 중요성이 날로 강조되고 있다. 그러나, 국내 많은 기관과 기업은 데이터 품질 저하로 인해 다양한 어려움을 겪는 중이다. 이 같은 상황에서 데이터 산업진흥 및 이용촉진에 관한 기본법(데이터산업법)이 지난 2021년 10월 제정됐고, 관련 시행령에 따라 2022년 4월 발효됐다.

해당 데이터산업법에는 데이터 품질관리와 데이터 산업 활성화를 위해 데이터품질인증기관, 가치평가기관, 관련 협회 등의 지정 및 운영에 관한 내용이 담겨 있다. 이를 근거로 2023년 7월 공모를 통해 3개의 데이터품질인증기관(TTA, CAS, 와이즈스톤)이 선정되면서 데이터품질인증(DQ(Data Quality)인증) 제도가 시행됐다. 이후 운영 규정 마련, 심사 체계 수립 등 약 3개월 간의 준비 과정을 거쳐, 당해 11월 각 DQ인증기관이 DQ인증 서비스를 본격적으로 개시했다. 2024년 8월엔 정형 데이터 내용에 국한해 시행하던 DQ인증 서비스가 비정형 데이터 내용, 데이터 관리체계 등으로 확대됐다.

이번 원고에선 DQ인증기관 중심으로 그간 마련한 DQ인증 심사체계를 소개하고, 정형 데이터 내용에 대한 DQ인증 사례를 살펴본다.

2. DQ인증 소개

2.1 품질기준

데이터산업법과 시행령에 따른 과학기술정보통신부「데이터 품질인증기관 지정 및 운영에 관한 지침」에선 <표 1>과 같은 10개 품질기준을 정의하고 있으며 이를 준용해 3절과 4절에서 소개하는 데이터 내용 및 데이터 관리체계에 대한 세부적인 심사 항목을 마련했다. 데이터 내용에 대한 인증은 완전성, 유효성, 일관성, 정확성, 접근성, 유일성 등 6개 품질기준에 따라 심사 항목을 도출했다. 데이터 관리체계에 대한 인증은 완전성, 유효성, 일관성, 정확성, 보안성, 유용성, 접근성, 적시성, 유일성 등 9개 품질기준과 관리체계 고유의 기본 품질기준 1개에 대응되는 심사 항목을 도출했다.

<표 1> DQ인증 품질기준

기준	내용
완전성	사용목적에 맞게 설계된 데이터가 필수항목을 누락 없이 포함해 완전한지 진단하는 기준
유효성	데이터가 규정된 형식과 의미, 관계 규칙을 준수하고 논리적으로 타당한지 진단하는 기준

일관성	준수해야 하는 표준, 규칙, 기준에 따라 데이터가 일관되게 입력돼 있는지 진단하는 기준
정확성	데이터가 업무규정 또는 논리에 따라 구조적, 의미적으로 정확한지 진단하는 기준
보안성	데이터가 보안 정책에 따라 안전하게 보호되고, 적절히 관리되고 있는지 진단하는 기준
유용성	데이터가 사용자의 데이터 활용 목적 및 요구사항에 부합하는 정보를 제공하는지 진단하는 기준
접근성	데이터가 쉽게 접근할 수 있고 사용하기에 편리한지 진단하는 기준
적시성	데이터가 최신성을 유지하고 사용자가 원하는 시점에 적합한 데이터를 적절하게 제공할 수 있는지 진단하는 기준
다양성	데이터가 사용자의 이용 목적에 적합한 다양한 특성을 포괄하고 편향되지 않았는지 진단하는 기준
유일성	데이터가 중복되지 않고 유일한지 진단하는 기준

2.2 인증 절차

DQ인증의 절차는 다음 세 단계로 구성된다.

인증신청:
신청인이 인증신청서 등을 통해 인증기관에 신청하고 신청 서류가 이상이 없으면, 신청인과 인증기관은 계약을 체결한다. 이후 상담 등을 통해 인증심사 및 심의에 필`한 사항을 확보한다.

↓

인증심사:
인증기관은 인증심사팀 구성 등 인증심사를 위한 계획을 수립한 후 본격적인 심사를 통해 품질 기준을 충족하는지 평가하고, 인증심사결과(평가)보고서를 작성해 인증심의위원회에 안건을 상정한다.

↓

인증심의:
인증심의위원회는 심사 결과를 심의해 인증 적합 여부를 최종적으로 결정하고, 인증 적합 시 인증기관은 신청기관에 인증서를 발급한다. 상기 인증심의위원회는 인증기관별로 데이터 관리 및 데이터품질 분야 학계와 산업계 등 외부 전문가들로 구성, 독립적이고 공정한 심의를 도모한다.

2.3 데이터 내용 인증 소개

데이터 내용 인증은 정형 데이터와 비정형 데이터로 구분해 심사체계를 구축한 관계로 각각의 특성에 따라 세부 심사 항목이 다르다. 한편, 인증 대상 데이터 유형은 데이터 복잡도와 심사 항목 수에 따라 Simple-Type, Normal-Type, Complex-Type 등 세 가지로 분류한다. 인증 등급은 Class A, Class B, Class C로 나뉘며, 결과에 따라 인증 거부가 발생할 수도 있다. 시간에 따라 변화하는 데이터 특성과 인증받는 기관의 입장을 고려하고 절충해, 인증 유효기간은 1년으로 설정했다.

2.4 데이터 관리체계 인증 소개

인증 등급은 Level 2, Level 3, Level 4, Level 5로 구분된다. 각 등급은 체계적이고 안전한 데이터 관리 시스템을 구축하는 데 필요한 요건 충족 수준을 반영한다. 관리체계의 특성을 고려해 인증 유효기간은 3년으로 설정했다.

이 밖에도 DQ인증에 대한 일반적 사항 상세 내용은 참고문헌 [9]의 1권을 참고한다.

3. 데이터 내용 인증의 심사체계

3.1 데이터 내용 심사 항목

정형 데이터 내용 심사 항목은 ISO/IEC 25024를 기반으로 일련의 부합화 과정을 거쳐 도출했으며, 그 주안점은 <표 2>와 같다.

<표 2> DQ인증 품질기준

순번	부합화 과정에서의 주안점
1	ISO/IEC 25024의 측정항목 중 'Target entities(측정 대상)'의 데이터 내용에 해당하지 않는 항목 제외(Contextual schema, conceptual data model, Data Models, Architecture, Document, form, presentation device)
2	데이터 내용의 심사 기준에 해당하는 항목 선정
3	ISO/IEC 25024 측정항목 중 'System Dependent' 관점으로 측정 불가한 항목 제외 (데이터 업데이트 지연 시간, 데이터값 추적성, 비 취약점, 사용자 접근 추적성, 데이터값 이해성, 연결된 마스터 데이터 이해성, 데이터 가용성 비율 등)
4	의미적으로 중복되는 심사 항목 제외
5	ISO/IEC 25024의 측정항목 중 사용 레벨*(level of use)이 REF인 항목은 제외 * ISO/IEC 25024의 Annex(부록) D에서 심사 항목 사용 레벨을 HR(Highly Recommendable), R(Recommendable), REF(for Reference) 등 세 단계로 구분하고 있으며, HR은 필수항목, R은 추천 항목, REF는 선택 항목으로 해석할 수 있음

한편 비정형 데이터 내용 심사 항목은 ISO/IEC 52 59-2 기반이나, 국내 단체표준인 TTAK.KO-10.1344-Part2(유통·활용 데이터 점검 방법 – 제2부: 비정형 데이터 품질지표)와 NIA(한국지능정보사회진흥원, National Information Society Agency)에서 발간한 '인공지능 학습용 데이터 품질관리 가이드 3.1'에서 정의하고 있는 비정형 데이터 심사 항목을 일부 채택해 도출했다.

이렇게 도출된 데이터 내용 심사 항목은 <표 3>과 같으며 항목별로 점수를 산정하는 일반적인 수식은 다음과 같다. 이 외에도 데이터 내용 심사 항목별 상세 설명과 수식은 참고문헌 [9]의 2권을 참고한다.

· X = 1 - (A / B)
· A = 해당 심사 항목의 기준에 따라 오류가 있는 데이터 아이템 개수
· B = 해당 심사 항목을 적용해 진단할 필요가 있는 전체 데이터 아이템 개수

3.2 데이터 유형 및 인증 등급

데이터 구조나 구성의 복잡도를 평가해 인증 대상 데이터를 <표 4>와 같이 분류하고 인증 등급은 <표 5>와 같이 구분해 인증을 부여하므로 인증서에는 이 두 가지 사항이 기재된다.

<표 3> 데이터 내용 심사 항목

구분	정형 데이터		비정형 데이터	
	품질기준	심사 항목	품질기준	심사 항목
필수 심사 항목	완전성(2개)	레코드 완전성, 데이터 값 완전성	완전성(3개)	데이터 파일 완전성, 레코드 완전성, 메타 데이터값 완전성
	유효성(4개)	구문 유효성, 의미 유효성, 범위 유효성, 관계 유효성	유효성(3개)	데이터 구조 구문 유효성, 데이터 포맷 유효성, 객체 유효성
	일관성(1개)	참조 무결 일관성	-	-
	-	-	정확성(2개)	구문 정확성, 메타 데이터 정확성
	-	-	유일성(1개	객체 유일성
선택 심사 항목	-	-	완전성(2개)	어노테이션 완전성, 특집 완전성
	유효성(1개)	데이터 값 정밀성	유효성(6개)	관계 유효성, 데이터 값 정밀성, 데이터 속성 유효성, 범위 유효성, 시간 유효성, 아노테이션 유효성
	일관성(2개)	공동 어휘 일관성, 데이터 포맷 일관성	일관성(2개)	공동 어휘 일관성, 데이터 포맷 일관성
	정확성(2개)	메타 데이터 정확성, 데이터 값 정확성	정확성(2개)	어노테이션 정확성, 주제 정확성
	접근성(1개)	표준기반 데이터 접근성	접근성(2개)	데이터 포맷 접근성, 표준기반 데이터 접근성
	유일성(1개)	데이터 값 유일성	유일성(1개)	레코드 유일성

<표 4> 인증 대상 데이터 유형의 구분

데이터 유형	기준	참고
Complex-Type	- 가능 필수 항목 모두 적용 - 선택 항목 3개 이상 적용	데이터 구조가 매우 복잡해 많은 심사 항목이 적용될 수 있는 유형으로 다양한 관계규칙, 업무규칙이 요구된다.
Normal-Type	- 가능 필수 항목 모두 적용 - 선택 항목 3개 미만 적용	중간 수준의 데이터 복잡도를 가져 상당수 심사 항목이 적용될 수 있는 유형으로 기본적인 관계, 업무규칙이 요구된다.
Simple-Type	- 필수 항목 일부 적용	데이터 구조가 단순하며, 필수 심사 항목 중 일부만 적용되는 유형으로 데이터의 관계와 조건이 단순하다.

<표 5> 인증 대상 데이터 유형의 구분

인증 등급	기준
Class A	개별 심사 항목 점수 0.95 이상, 심사 항목 전체 평균 점수 0.99 이상
Class B	개별 심사 항목 점수 0.95 이상, 심사 항목 전체 평균 점수 0.97 이상
Class C	개별 심사 항목 점수 0.95 이상, 심사 항목 전체 평균 점수 0.95 이상

4. 데이터 관리체계 인증의 심사 체계

데이터 관리체계 심사 항목은 ISO 8000-61에서 정의하고 있는 프로세스를 기반으로 도출했으며, 이와 관련해 데이터 관리체계 심사 항목을 '프로세스'라고도 칭한다. 인증 등급은 ISO 8000-62의 성숙도 수준을 준용해 총 5개로 구분되며, 데이터 관리체계에 대해 'Level 2'부터 인증을 부여하

므로 'Level 1'로 판정되면 인증을 부여하지 않는다.

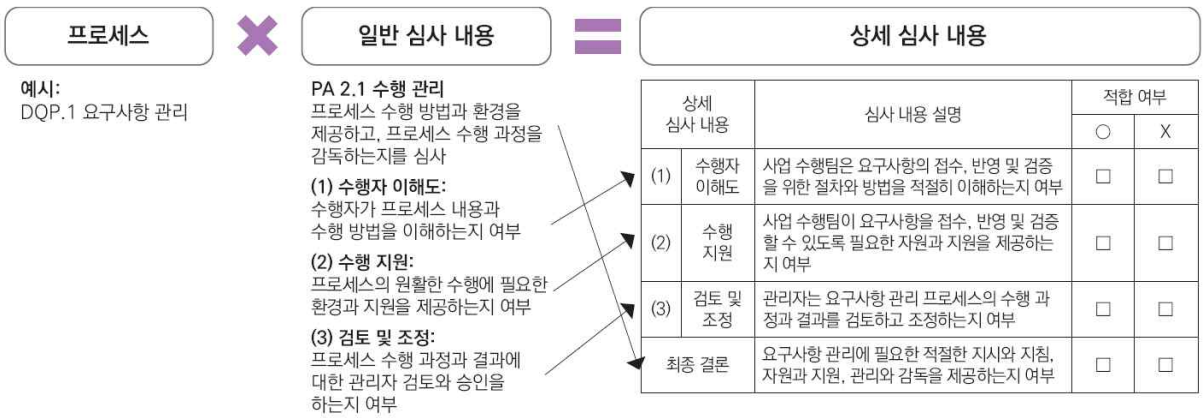
또한, ISO 8000-62에선 프로세스별로 요구되는 속성을 정의하고 있으며 ISO/IEC 33020에는 프로세스 속성별 적/부 여부를 판별하기 위한 상위 수준 점검 목록을 정의하고 있다. 데이터 관리체계 인증 심사체계에선 이를 부합화해 점검 목록을 도출했다.

<표 6>은 데이터 관리체계의 인증 등급별 심사 대상이 되는 프로세스, 요구되는 프로세스의 속성, 인증 부여 여부를 나타낸다. 여기서, 높은 숫자의 레벨은 낮은 숫자의 레벨 심사 대상 프로세스와 요구 속성을 포함한다. 이 외에도, 부합화한 점검 목록과 세부점검 목록은 참고문헌 [9]의 3권을 참고한다.

<표 6> 인증 등급별 심사 대상 프로세스, 프로세스 속성, 인증 부여 여부

인증 등급	심사 대상 프로세스	프로세스 속성	인증 여부
Level 1	데이터 처리, 데이터 보안 관리	프로세스 수행	인증 미부여
Level 2	요구사항 관리, 데이터 명세 및 작업지시서 제공, 데이터 품질 모니터링 및 통제	성과 관리 작업 산출물 관리	인증 부여
Level 3	데이터 품질 전략 관리, 데이터 품질 실행 계획, 데이터 품질 정책/표준/절차 관리, 데이터 아키텍처 관리, 데이터 품질 조직 관리 데이터 운용 관리, 데이터 클렌징	프로세스 정의 프로세스 전개	
Level 4	데이터 품질 이슈 점검, 측정 기준 제공, 측정 결과 평가 데이터 품질 프로세스 성과 측정 데이터 흐름 관리, 인적 자원 관리	정량적 분석 정량적 통제	
Level 5	근본 원인 분석 및 해결 방안 개발 데이터 오류 방지 위한 프로세스 개선	프로세스 혁신 프로세스 혁신 구현	

한편 데이터 관리체계 인증심사 시엔 인증 의뢰 기관이 요청한 Level 수준의 프로세스와 속성에 대해, 상위 및 세부 점검 목록을 바탕으로 서류 심사와 현장 심사를 병행한다. [그림 1]은 특정 프로세스(요구사항 관리)에 대한 인증심사 과정의 예를 나타낸다.



[그림 1] 특정 프로세스 인증심사 과정

5. 정형 데이터 내용 인증 사례

5.1 심사 대상 데이터

이번 사례의 심사 대상 데이터는 암 환자에게 병기 별 맞춤형 암 정보 콘텐츠(논문, 칼럼, 뉴스, 유튜브 영상, FAQ 등), 환자-의사 간 질의응답(QnA) 서비스를 제공하기 위한 것이다. 암 환자 질병(암 종류), 병기에 맞는 표준치료 계획, 치료 과정에서의 부작용, 주의 사항 등의 정보를 관계형 데이터베이스에 구축했으며 주요 명세는 아래와 같다.

- 데이터 유형: 정형 데이터
- 데이터 형식: 관계형 데이터베이스(RDBMS)
- 테이블 수(심사 대상/전체): 29/30
- 컬럼 수(심사 대상/전체): 278/288

여기서 특정 1종(fcm_queue)의 테이블은 푸시(push) 메시지를 잠시 보관(queue)하기 위한 임시 테이블로, 인증의뢰기관과 인증기관 간 협의를 통해 심사 대상에서 제외했다.

5.2 심사 결과 및 심사 항목 판정의 예

심사 결과는 <표 7>와 같이 요약되며 필수 심사 항목 전체 통과(각 항목 0.95 이상)되고, 전체 평균 0.99 이상으로 심사 체계에 따라 Normal-Type 유형 Class A 등급으로 판정됐으며, 추후 심의위원회를 거쳐 인증이 부여됐다. 한편, 심사 항목 중 의미 유효성의 상세 심사 결과를 살펴보면, 전체 컬럼 중 유효한 값 목록을 갖는(코드 정의서 존재) 컬럼을 대상으로 정의된 코드만 사용해 데이터 값이 입력돼 있는지 점검했고, 특정 컬럼에서 코드로 존재할 수 없는 값 117건이 오류로 검출됐다.

6. 맺음말

TTA는 지난 2023년 11월 인증 서비스 개시 이후 총 11건 인증을 부여했으며, 이 중 6건은 정형 데이터 내용, 3건은 비정형 데이터 내용, 2건은 데이터 관리체계에 대한 인증이다. 특히 데이터 관리체계에 대한 인증은 TTA가 국내 최초로 지난 2024년 10월 부여했다.

앞으로 TTA를 비롯한 DQ인증기관은 더 폭넓은 분야 데이터와 관리체계를 대상으로 인증 서비스를 적극적으로 확대해 나갈 예정이며, 인증제도 활성화와 저변 확대를 촉진하기 위해 홍보 활동과 설명회 개최 등 다양한 노력을 이어갈 계획이다.

<표 7> 인증 사례 심사 결과

결과 요약						
심사항목(필수)	세부 진단항목	진단대상 구분	총진단건수 (①)	오류건수 (②)	오류율 (②/①)	점수 (1-②/①)
1. 데이터 값 완전성	필수	진단대상 컬럼 : 43 오류발생 컬럼 : 0	21,651	0	0	1.00000
2. 데이터 레코드 완전성	-	진단대상 레코드 : 10,097 오류발생 레코드 : 0	10,097	0	0	1.00000
3. 구문 유효성	구문(날짜)	진단대상 레코드 : 38 오류발생 레코드 : 0	8,759	0	-	-
	구문(번호)	진단대상 컬럼 : 3 오류발생 컬럼 : 2	49	20	-	-
	소계	진단대상 컬럼 : 41 오류발생 컬럼 : 2	8,808	20	0.002271	0.99773
4. 의미 유효성	코드	진단대상 컬럼 : 28 오류발생 컬럼 : 1	15,175	117	0.007710	0.99229
5. 범위 유효성	수량	진단대상 컬럼 : 10 오류발생 컬럼 : 0	2,735	0	0	1.00000
6. 관계 유효성	논리관계	진단대상 테이블 : 2 오류발생 테이블 : 0	437	0	0	1.00000
7. 참조 무결 일관성	참조	진단대상 테이블 : 17 오류발생 테이블 : 0	10,829	0	0	1.00000
평균					0.001426	0.99857

[참고문헌]

[1] ISO/IEC 25024:2015. 'Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality.'

[2] 인공지능 학습용 데이터 품질관리 가이드라인 v3.1, 한국지능정보사회진흥원, 2024년 1월

[3] TTAK.KO-10.1344-Part2 유통·활용 데이터 점검 방법 – 제2부: 비정형 데이터 품질지표, 2023년 12월

[4] ISO/IEC 5259-2(en), Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

[5] ISO 8000-61 Data quality management: Process reference model

[6] ISO 8000-62 Data quality management: Organizational process maturity assessment: Application of standards relating to process assessment

[7] ISO/IEC 17021-1 Conformity assessment — Requirements for bodies providing audit and certification of management systems

[8] ISO/IEC 33020 Information technology — Process assessment — Process measurement framework for assessment of process capability

[9] DQ인증 가이드라인 1권~3권, 2025년 2월, 과학기술정보통신부 및 한국데이터산업진흥원

[주요 용어 풀이]

- 정형 데이터(Structured Data): 구조화돼 체계적으로 정리된 형태의 데이터. 스프레드시트 파일, 관계형 데이터베이스처럼 명확한 스키마와 고정된 필드 형식을 갖는다.
- 비정형 데이터(Unstructured Data): 이미지, 비디오, 사운드, 텍스트 등 구조나 형식 없이 저장되는 데이터
- 반정형 데이터(Semi-structured Data): 키와 값의 계층 구조로 구성되는 반구조화된 데이터. XML, JSON 등의 파일 형식으로 저장된다
- 데이터 항목(Data Item): 식별 가능한 가장 작은 데이터 단위(데이터의 '컨테이너'로 볼 수 있고, 필드는 데이터 항목과 동의어로 간주됨)
- 데이터 레코드(Data Record): 단위로 처리되는 관련 데이터 항목의 집합(표 형태 데이터에서 한 행에 해당. 비정형 데이터의 경우, 단위 비정형 데이터와 상응 라벨 데이터 쌍으로 간주됨)
- 메타 데이터(Meta Data): 데이터를 설명하는 데이터(정형 데이터의 경우, 데이터베이스의 구조와 정의, 사용자 정보 등에 관한 데이터를 의미하고, 비정형 데이터의 경우, 단위 비정형 데이터와 그에 대한 메타 데이터 또는 라벨링 데이터를 데이터 레코드로 볼 수 있음)

※ 출처: TTA 저널 제218호