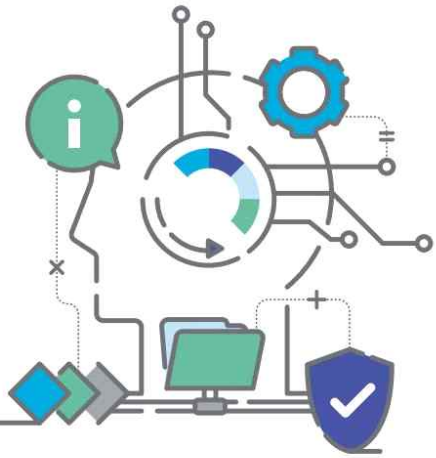
2권
데이터품질평가 종합안내서
데이터 거래 지원 가이드라인

Kdata 한국데이터산업진흥원
Korea Data Agency

※ 본 가이드는 데이터 유통 및 활용 촉진을 지원하기 위한 연구 결과물로서, 데이터 거래 시 참고할 수 있도록 가격·품질·법적 쟁점 등의 이해를 돕기 위한 자료입니다.

2권. 데이터품질평가 종합안내서

CONTENTS



PART 01

Korea Data Agency

추진배경 및 목적

1.1. 추진배경

4차 산업혁명에 따라 데이터는 사람, 자본 등 기존의 생산요소를 능가하는 핵심 자원으로 부상하며 전체 산업의 혁신성장을 가속화하고 있다. 시장조사업체 IDC에 따르면 세계적으로 발생하는 데이터량은 2016년 16제타바이트(ZB)에서 2025년 163ZB로 10배 이상 증가할 전망이다.

데이터를 '21세기 원유'로 표현하듯 데이터는 원유와 같이 활용 목적에 맞게 제대로 정제하고 유통해야 비로소 가치 있는 데이터상품으로 탄생한다. 데이터의 수집과 저장, 유통, 활용 등 생태계의 가치사슬을 기반으로 공급-중개-수요를 연결하는 데이터시장이 탄탄하게 구축되어야 비로소 경제적 가치를 창출할 수 있다.

정부가 데이터스토어와 데이터바우처, 빅데이터 플랫폼, 마이데이터 등의 사업을 통해 데이터 거래를 촉진하고 있지만 현재로서는 양질의 데이터가 부족하고 데이터품질이나 가격 문제, 개인정보보호 등 법·제도 규제 등이 거래 활성화를 가로막는 걸림돌로 파악되고 있다.
데이터품질을 4차산업의 핵심적인 요소로 생각하고 이를 향상시키려는 활동이 일상적으로 이루어져야 한다. 4차 산업혁명으로 인해 데이터 활용이 많아질수록 데이터품질에 대한 요구는 증가한다.

현재 국내 데이터품질 활동은 정형데이터 위주의 표준, 구조, 값, 관리체계 중심으로 연구·활용되고 있으며 비정형데이터의 경우 정성적 관점의 실측으로 평가되고 있다.
또한 데이터스토어와 데이터거래소, 빅데이터 플랫폼 등의 다양한 데이터 거래 플랫폼에서 서로 다른 데이터품질 기준과 지표를 적용하여 데이터상품을 평가함으로써 객관성과 신뢰성이 저하되는 현실이다.

이러한 문제를 해결하기 위해서 데이터거래소 품질관리의 객관성과 일관성 확보 차원에서

종합 안내서를 제시함으로서 데이터거래소 전반의 데이터품질 수준을 향상 시켜야 할 중요한 시기이다.

본 종합 안내서는 이런 필요성을 바탕으로 국외(미국, 영국, 호주, 캐나다, 독일 등) 데이터 품질 관련 선진 사례와 국내 사례 조사를 통해 다양한 시사점을 도출하여 반영하였다.
현실적인 종합안내서 수립을 위해 실제 데이터거래소의 데이터상품을 대상으로 품질 실태 조사를 실시하였으며, 조사 결과를 바탕으로 데이터상품의 유형과 문제점을 분석하였다.

- 데이터크롤링을 통한 실제 데이터상품의 데이터품질 진단 시뮬레이션 수행
- 설문을 통한 품질관리 수준 조사 실시

또한 현재 정부·민간에서 활용중인 데이터품질 관련 타 지침과의 관계분석을 통해 데이터 거래소의 데이터상품 품질제고를 위해서 꼭 필요한 부분에 대한 통합과 상호연계를 고려하였다.

1.2. 추진목적

본 안내서는 데이터거래시장 참여자가 공통된 시각으로 데이터상품의 품질을 평가할 수 있는 데이터품질평가 모델을 제안한다.

- 비정형 데이터상품의 육안 검사는 데이터품질 평가에 상당한 시간과 인력 투입이 불가피하다. 대규모 예산과 인력이 확보된 대기업이나 중앙기관이 아니라면 비정형 데이터상품에 대한 품질관리는 쉽게 접근할 수 없다. 따라서 기계적 데이터 품질평가방법의 모형을 개발하여 판매자(공급자), 데이터 거래소, 수요자(구매자) 모두가 활용 가능한 품질평가 지표를 제공한다.
- ISO8000, ISO/IEC 25024, 빅데이터플랫폼 및 센터 데이터 품질관리 가이드, 가공데이터 품질 가이드라인, 공공정보 데이터 품질관리 매뉴얼등을 통해 다양한 품질지표가 제정되고 활용되고 있다. 본 안내서에서는 기존 데이터 품질평가지표의 장단점을 분석하여 고도화하고 데이터상품의 특성을 고려하여 객관성과 활용성이 확보된 품질평가지표를 신규로 제정한다.
- 품질평가모형의 객관성과 활용성 확보를 위해 정성적 평가 및 육안평가 방법은 최대한 배제하고, 시스템화하여 정량적 측정이 가능한 지표들로 구성한다.
- 품질평가의 기초가 되는 메타데이터와 연계하여 평가할 수 있는 품질평가지표를 개발하고 품질평가 결과는 가치평가 모델에 제공하여 활용될 수 있도록 한다.

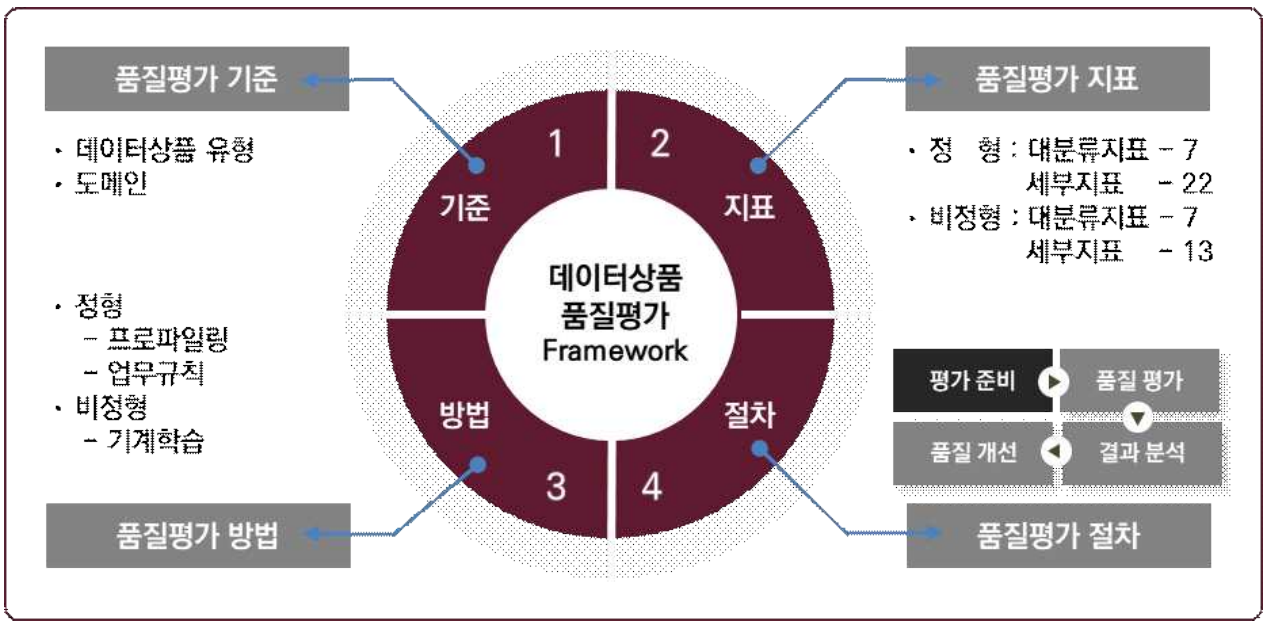
PART 02

Korea Data Agency

품질평가 프레임워크

2.1. 개요

품질평가 프레임워크란 데이터상품의 서비스 수준 제고를 위해 데이터상품 공급자, 구매자, 거래소 차원에서 데이터상품 품질평가에 대한 기준을 정립하고, 기존 데이터의 품질평가·개선 절차 및 방법 등을 체계화한 것을 말한다.



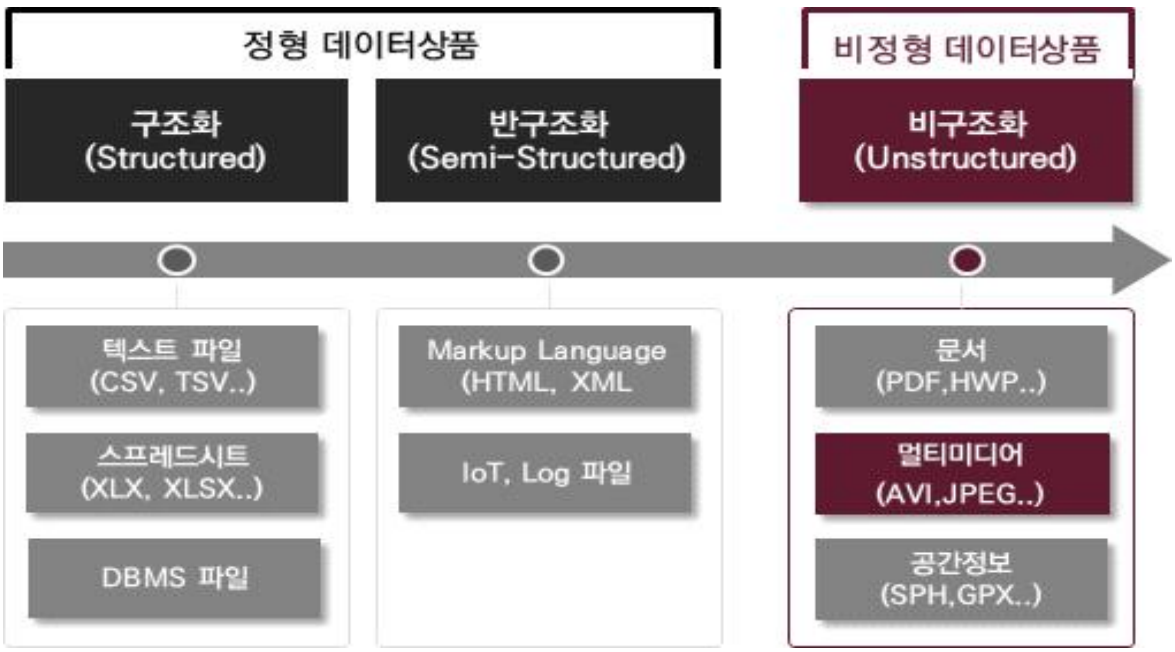
[그림 2-1] 데이터상품 품질평가 프레임워크

2.2. 적용 대상

- 데이터상품 품질평가 프레임워크 적용 대상은 데이터상품을 생산하는 공급자(생산·가공), 데이터상품을 거래하는 거래소, 데이터상품을 활용하는 구매자이며 해당 기관 및 기업 내 모든 조직이 프레임워크의 활용 대상이다.
- 데이터거래소는 데이터상품의 규모에 따라 품질평가를 위한 단계별 절차와 산출물을 필수 요소와 선택 요소를 구분하여 상황에 맞게 사용할 수 있다. 또한 가용자원을 적절하게 분배하여 품질 제고에 노력하여야 한다.
- 거래소를 통해 유통되는 데이터상품을 대상으로 하며, 데이터상품의 유형을 분류하면, 크게 "정형 데이터"와 "비정형 데이터"로 구분된다.

  정형데이터는 CSV, TSV, 스프레드시트등 파일로 제공되는 "구조화 데이터(Structured Data)"와 JSON, HTML, XML등 API로 제공되는 "반구조화 데이터(Semi-Structured Data)"로 분류한다.

  비정형 데이터는 문서, 멀티미디어, 공간정보 지도, IoT, 소셜네트워크 데이터 등의 비구조화 데이터(Unstructured Data)를 의미한다.



[그림 2-2] 적용 대상 데이터상품 유형

2.3. 적용 범위

- 측정 가능한 데이터상품의 데이터 값과 메타데이터
- 시스템적으로 측정이 불가능한 평가지표는 대상에서 제외
- 비정형데이터는 이미지 데이터상품을 중심으로 시스템으로 측정 가능한 지표를 도출

2.4. 가이드 특징

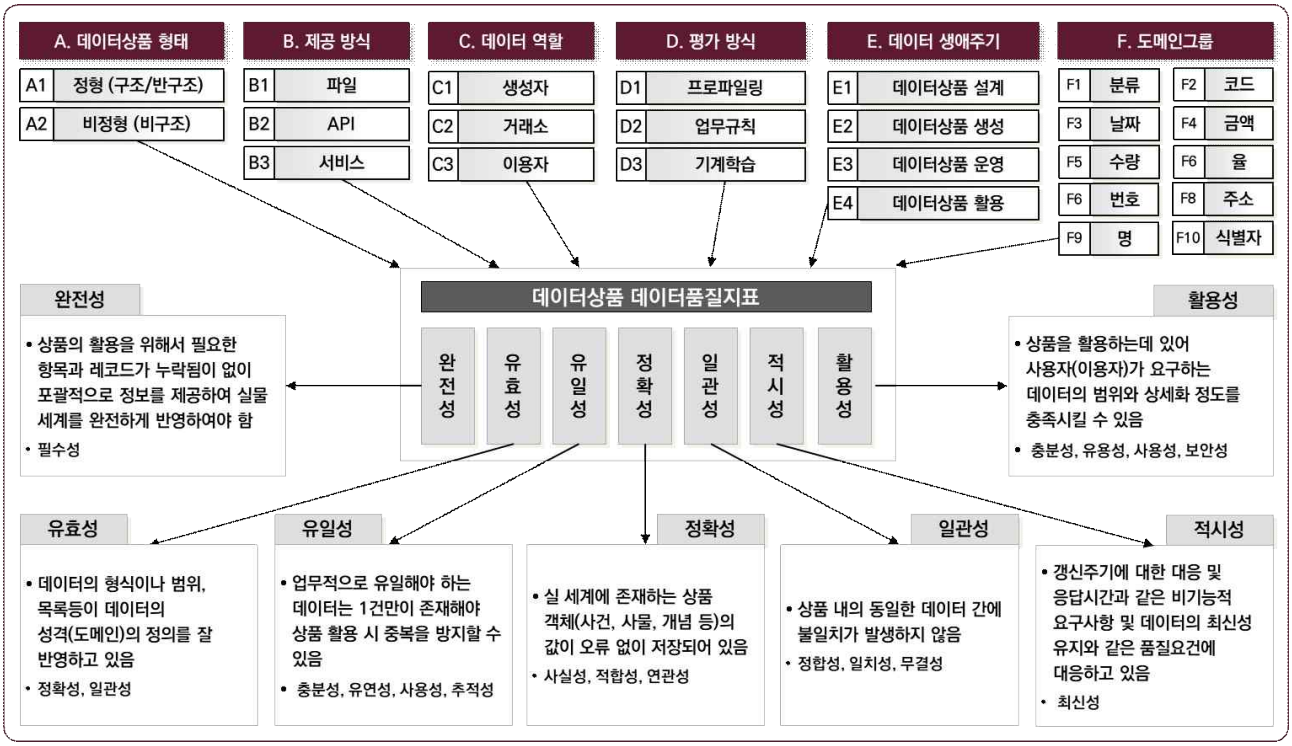
- 비정형 데이터에 대해 품질평가지표를 시스템화하여 정량적으로 평가할 수 있는 기준을 제공한다.
- 데이터상품 특성에 따른 평가 방법을 제공하여 본 가이드에서 제공한 품질평가지표를 기반으로 품질평가 시스템으로 구현 가능하고 정량화가 가능하다.
- 데이터상품 특성을 고려한 데이터 품질평가 방법으로 데이터거래소의 데이터 품질관리 객관성과 활용성을 높일 수 있다.
- 데이터상품 특성을 반영한 품질평가지표, 품질평가 방법·기능을 활용관점에서 이해하기 쉽게 제공하여 품질평가 활용 및 도입의 효과성이 높다.
- 데이터상품 생성, 제공, 활용 시 품질평가 절차와 그에 따른 품질평가를 정의하고 있어 현재 운영 중인 거래소의 데이터상품에 대해서도 품질평가 활동이 가능하다.

PART 03

Korea Data Agency

품질평가 기준

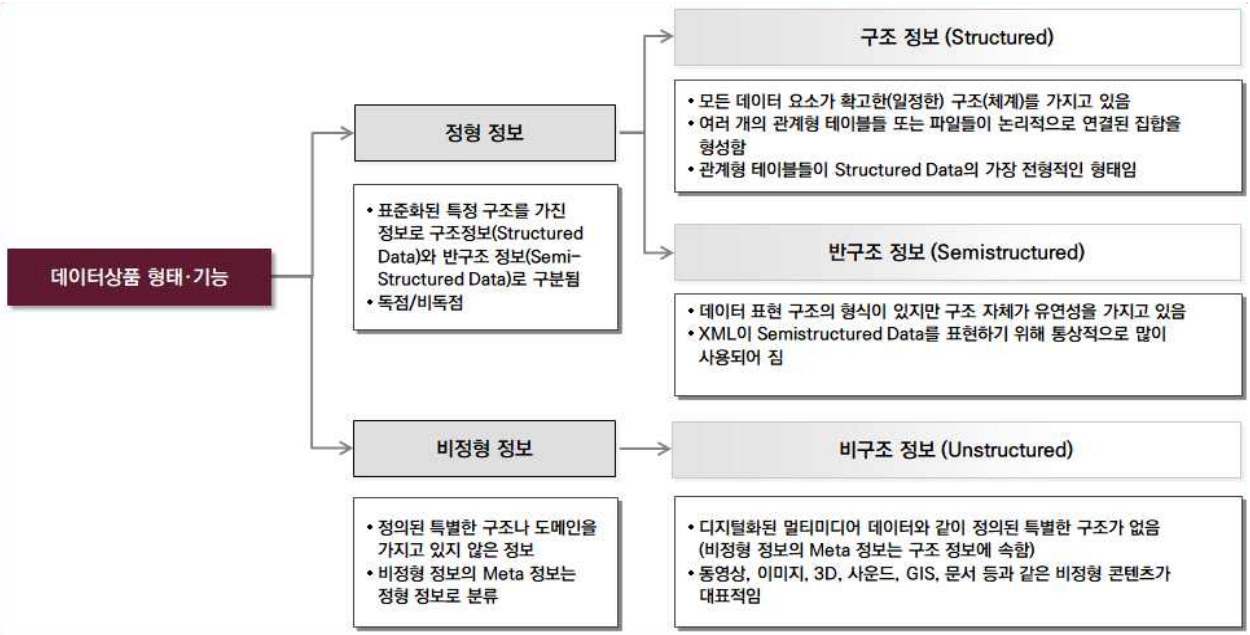
데이터상품의 품질평가 기준은 “형태/유형”, “제공방식” , “역할/권한”, “평가형식”, “생애주기”, “도메인”을 정의하고 이를 바탕으로 데이터품질평가지표가 구성된다.



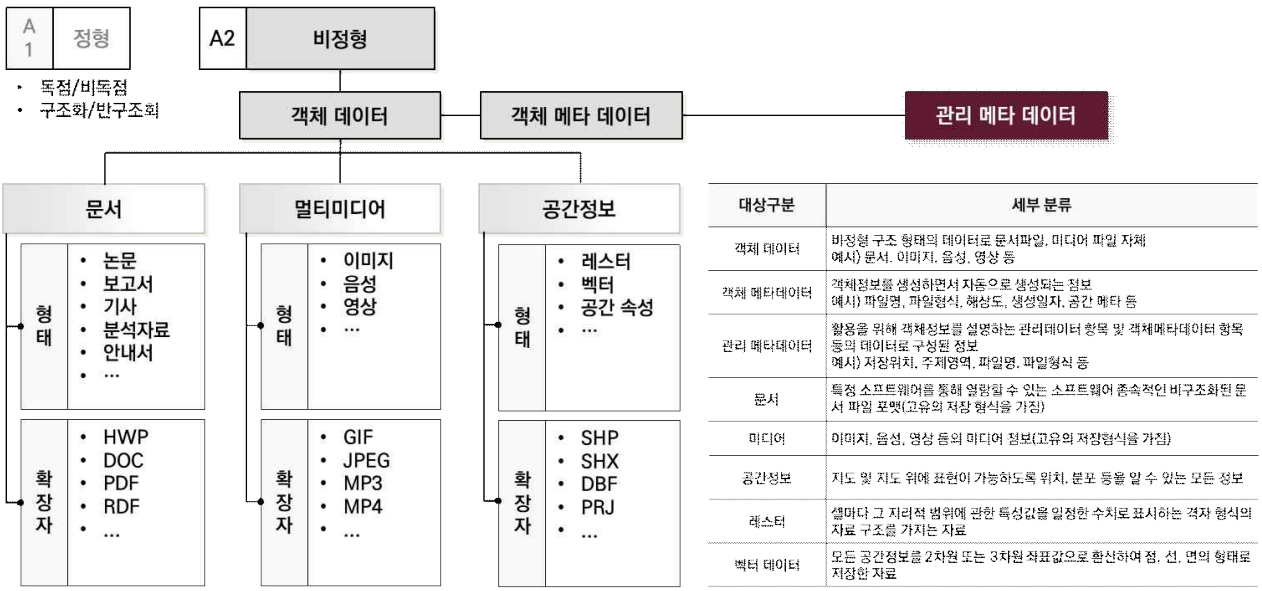
[그림 2-3] 데이터상품 품질평가 프레임워크 구성요소

3.1. 데이터상품 형태

데이터상품의 데이터상품 형태는 품질평가 방법의 적용 용이성을 기준으로 한 분류인 정형/비정형 분류하고 다음과 같이 정의한다.



[그림 2-4] 데이터상품 형태분류



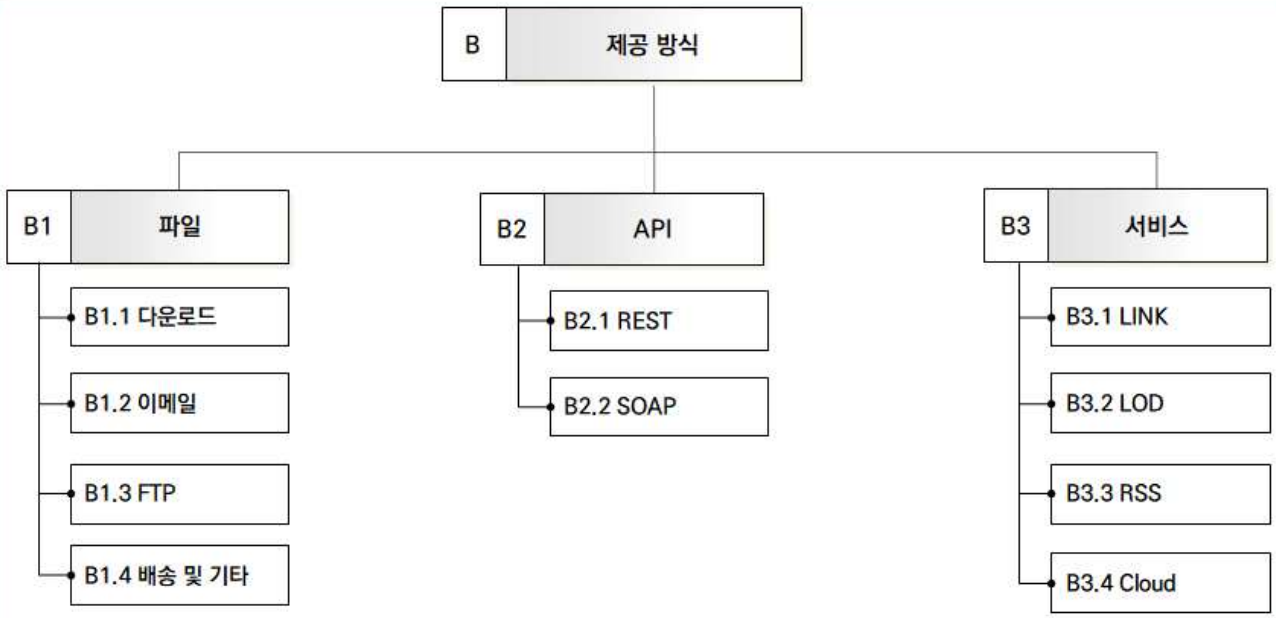
대상구분	세부 분류
객체 데이터	비정형 구조 형태의 데이터로 문서파일, 미디어 파일 자체 예시) 문서, 이미지, 음성, 영상 등
객체 메타데이터	객체정보를 생성하면서 자동으로 생성되는 정보 예시) 파일명, 파일형식, 해상도, 생성일자, 공간 메타 등
관리 메타데이터	활용을 위해 객체정보를 설명하는 관리데이터 항목 및 객체메타데이터 항목 등의 데이터로 구성된 정보 예시) 저장위치, 주제영역, 파일명, 파일형식 등
문서	특정 소프트웨어를 통해 열람할 수 있는 소프트웨어 종속적인 비구조화된 문서 파일 포맷(고유의 저장 형식을 가짐)
미디어	이미지, 음성, 영상 등의 미디어 정보(고유의 저장형식을 가짐)
공간정보	지도 및 지도 위에 표현이 가능하도록 위치, 분포 등을 알 수 있는 모든 정보
레스터	셀마다 그 지리적 범위에 관한 특성값을 일정한 수치로 표시하는 격자 형식의 자료 구조를 가지는 자료
벡터 데이터	모든 공간정보를 2차원 또는 3차원 좌표값으로 환산하여 점, 선, 면의 형태로 저장한 자료

[그림 2-5] 데이터상품 형태 세부분류

3.2. 데이터상품 제공방식

품질평가지표는 데이터상품의 제공방식에 따라 상이하게 적용될 수 있으므로 국내·외 데이터 거래소의 데이터 제공방식을 조사하여 품질관점에서의 특징을 다음과 같이 정의한다.

- 가장 기본이 되는 “파일다운로드”, “API” 는 대부분 거래소에서 기본 제공방식으로 제공되고 있음
- REST, SOAP, WEB Service 등은 API로 분류됨
- LINK : 거래소외 공급자가 제공하는 페이지에서 별도 다운로드
- LOD, 메일, 배송등은 별도의 제공방식으로 분류



[그림 2-6] 데이터상품 제공방식 분류

3.3. 품질평가 점수화

품질평가 결과를 점수화하기 위해서는 정량적 진단의 프로파일링과 업무규칙등에 대한 품질평가지표별 진단 후 전체 건수 대비 오류건수에 대한 오류율(%), 정합성율(%)을 산정하고, 6시그마 관점에서 DPMO(EPM), 6시그마지수, 품질지수를 산출하여 수준 평가를 수행한다.

[표 2-1] 품질지수 산정기준

항 목	내 용
오류율	○ 계산방법 : 오류건수/전체건수 * 100
정합성율	○ 계산방법 : 100 - (오류건수/전체건수 * 100)
DPMO	○ 백만기회당 결함수 (Defects Per Million Opportunities) : 단위당 결함 수에서 결함이 발생될 수 있는 기회를 나타낸 결함수 ○ 계산방법 : 오류건수/전체건수 * 1,000,000
시그마(σ) 수준	○ DPMO < 0.3 이면 Sigma 값은 6.9이다 ○ DPMO > 933,192이면 Sigma값은 0 이다 ○ 0 < DPMO <933,192라면 Sigma의 값은 ABS(NORMSINV(소계/1000000) - 1.5) 이다 ○ NORMSINV : 표준정규누적분포의 역함수의 값. 평균은 0이고 평균편차는 1이다.
품질지수	○ Sigma > 6 이면 품질지수는 100이다 ○ 1.5 < Sigma < 6이면 품질지수는 (Sigma - 1.5)* 49.9/4.5 + 50이다 ○ Sigma <1.5이면 품질지수는 Sigma * 50/1.5

3.4. 품질평가 도메인

데이터상품에서의 도메인이란 데이터상품 개별 항목들이 설계 단계에서 부여되는 고유한 성격으로 데이터상품에서 관리하는 데이터의 가장 작은 단위의 항목에 대한 정의라고 할 수 있다.

도메인 분석은 데이터상품이 보유하고 있는 항목 정보를 기초로 도메인을 분류하고 정의하는 방법이다. 도메인 분석을 통해 해당 항목의 도메인을 정의하여 관리하면 도메인 특성에 부합된 데이터는 항상 무결성을 유지할 수 있다.

[그림 2-7] 도메인 분류

[표 2-2] 도메인 그룹

도메인그룹	도메인그룹 설명
명칭	○ 사람이나 사물 또는 무형의 행위, 상태 등을 표현하는 이름을 표현하기 위한 도메인그룹 *(예) 명, 성명 등*
내용	○ 서술 형식으로 상세 내용을 표현한 것으로 자유 형식의 텍스트를 표현하기 위한 도메인그룹 *(예) 내용, 내역, 비고, 사항, 설명 등*
주소	○ 사람 또는 기관, 회사가 자리잡고 있는 행정구역의 전체주소, 우편번호주소, 상세주소 및 인터넷상에서 연결된 컴퓨터나, 웹 페이지를 찾아가는 주소를 표현하기 위한 도메인그룹 *(예) 주소, 우편번호주소, 상세주소, IP주소, 이메일주소 등*
수량	○ 객체의 개수나 량을 수로써 표현하기 위한 도메인그룹으로, 일반적인 측량단위(평수 등)도 포함됨. *(예) 건수, 수량, 점수, 년수, 좌수, 일수, 개월수 등*
금액	○ 돈의 가치를 수로 표현하는 도메인그룹으로, 금액의 손익(Profit) / 금액의 수익(Earnings) / 금액의 비용(Cost)관련 가치를 표현 *(예) 금액, 손익(Profit), 수익(Earnings), 비용(Cost), 가격, 요금, 세금 등*
율	○ 이율 / 확률 / 환율 / 지수 등의 비율을 수로 표현하기 위한 도메인그룹 *(예) 금리, 이율, 환율, 세율, 지수 등*
날짜	○ 'YYYYMMDDHH24MISS' 형식의 날짜데이터를 용도에 따라 일자, 시분초 등의 단위로 구분하여 표현하기 위한 도메인그룹 *(예) 일자, 일시, 년도 등*
번호	○ 수치형의 오름차순 번호 및 문자/숫자/기호를 포함한 번호체계를 표현하기 위한 도메인그룹 *(예) 순번, 일련번호, 사원번호, 전화번호, 휴대전화번호 등*
분류/코드	○ 여부 / 유무 및 데이터상품에서 사용되는 공통코드 및 개별코드를 표현하기 위한 도메인그룹 *(예) 여부, 유무, 금융회사코드 등*

[표 2-3] 도메인 분류

도메인그룹	도메인 예시	평가지표 / 평가내용
명칭	○ 명, ID, 장소, 고객명, 제목, 영문고객명	○ 유효성 - 포맷유효성 ○ 정확성 - 의미정확성
내용	○ 내용, 비고, 설명, 정보, 요약	○ 유효성 - 포맷유효성 ○ 정확성 - 의미정확성
주소	○ IP, 도로명, 법정동, 행정동, 이메일	○ 유효성 - 포맷유효성 ○ 유효성 - 목록유효성 ○ 일관성 - 참조무결성
수량	○ 건수, 매수, 회차, 개수, 거리, ○ 규모, 길이, 무게, 속도, 횟수, ○ 평형, 면적, 온도	○ 유효성 - 범위유효성 ○ 유효성 - 포맷유효성
금액	○ 금액, 세금, 가격, 단가, 비용, ○ 요금, 잔액, 총액	○ 유효성 - 범위유효성 ○ 유효성 - 포맷유효성
율	○ 금리, 이율, 비율, 환율, 백분율	○ 유효성 - 범위유효성 ○ 유효성 - 포맷유효성
날짜	○ 년월, 년, 년월일, 시, 분, 초, 일, 월, 월일 ○ 반기, 분기	○ 유효성 - 포맷유효성
번호	○ 주민등록번호, 사업자등록번호, ○ 우편번호, 고객번호, 계좌번호	○ 유효성 - 포맷유효성 ○ 번호의 패턴 및 채번 규칙을 평가
분류/코드	○ 개별코드, 통합코드	○ 유효성 - 목록유효성 ○ 일관성 - 참조무결성 ○ 일관성 - 항목 값 일관성

상품명세서의 컬럼정의 항목별로 도메인을 지정해야 하며 도메인에 따른 유효 기준을 다음과 같이 기술해야 한다.

[표 2-4] 도메인 유효 기준 정의

도메인그룹	유효 기준 정의 방법
명칭 내용	○ 입력될 수 있는 유효한 언어 또는 입력될 수 없는 언어를 정의한다. *(예) (형식 유효성) 한글, 영문대소문자, 특수문자만 입력가능* *(예) (형식 유효성) 비완성형 한글 입력 불가능* *(예) (형식 유효성) [ㄱ-ㅎ|ㅏ-|]+$* ○ 무의미한 단어로만 채워질 수 없음에 대해 정의한다. *(예) (의미 정확성) 동일한 하나의 단어만 구성 불가능, 2글자 이상*
주소	○ 주소의 참조값에 대해 정의한다. *(예) (참조무결성) 행정동주소 입력가능* ○ 주소의 생성규칙에 대해 정의한다. *(예) (형식 유효성) [0-9a-zA-Z]([-_.]?[0-9a-zA-Z])*@[0-9a-zA-Z]([-_.]?[0-9a-zA-Z])*.[a-zA-Z]*
수량 금액 율	○ 수치데이터의 범위에 대해 정의한다. *(예) (범위 유효성) 0보다 크고 999보다 작다* *(예) (범위 유효성) 컬럼 >= 0 AND 컬럼 <= 1000* ○ 수치데이터의 형식에 대해 정의한다. *(예) (형식 유효성) 숫자로만 구성되어야 한다.* *(예) (형식 유효성) [0-9]+$*
날짜	○ 날짜데이터의 유효 범위에 대해 정의한다. *(예) (범위 유효성) 컬럼 >= 19000101 AND 컬럼 <= 29991231* ○ 날짜데이터의 유효 형식 대해 정의한다. *(예) (형식 유효성) YYYYMMDD*
번호	○ 번호데이터의 유효 형식 대해 정의한다. *(예) (형식 유효성) 숫자3 - 숫자3* *(예) (형식 유효성) 999 - 999* *(범례) 숫자치환 : 9 , 한글치환 : H , 영문대문자 : A , 영문소문자 : a* ○ 범호데이터의 유효 채번규칙 대해 정의한다. *(예) (형식 유효성) 일련번호코드(3자리) + 개인 법인 구분코드(2자리) + 일련번호코드(4자리) +검증번호(1자리)*
분류/코드	○ 코드의 참조에 대해 정의한다. *(예) (목록 유효성) 1,2,3*

3.5. 품질평가지표

(1) 개요

품질평가지표(DQI - Data Quality Index, Data Quality Indicator)란?
데이터상품의 품질을 평가하기 위한 기준으로서 데이터상품의 활용을 위한 데이터의 신뢰를 충족시키기 위한 조건이며, 데이터상품의 결함을 최소화하기 위해 지속적인 품질진단을 통해 관리되어야 할 평가기준이다.
품질평가지표는 다음의 원칙에 따라 관리되어야 한다.

① 데이터 품질평가지표는 데이터상품의 품질을 측정하는 기준이므로 지속적으로 관리해야 한다.
② 데이터 품질평가지표는 특정 상품만을 위해 측정되는 지표가 아닌 전체 상품에 공통으로 적용되어야 한다.
③ 데이터 품질평가지표는 변동의 여지가 적어야 한다.
④ 데이터상품 품질 측정을 위해 선정된 데이터 품질평가지표는 공급자, 거래소, 구매자 사이에 공유 되어야 한다.
⑤ 반드시 시스템으로 정량적 측정이 가능해야 한다.
⑥ 측정할 수 없는 것은 관리할 수 없고 관리할 수 없는 것은 개선할 수 없다.

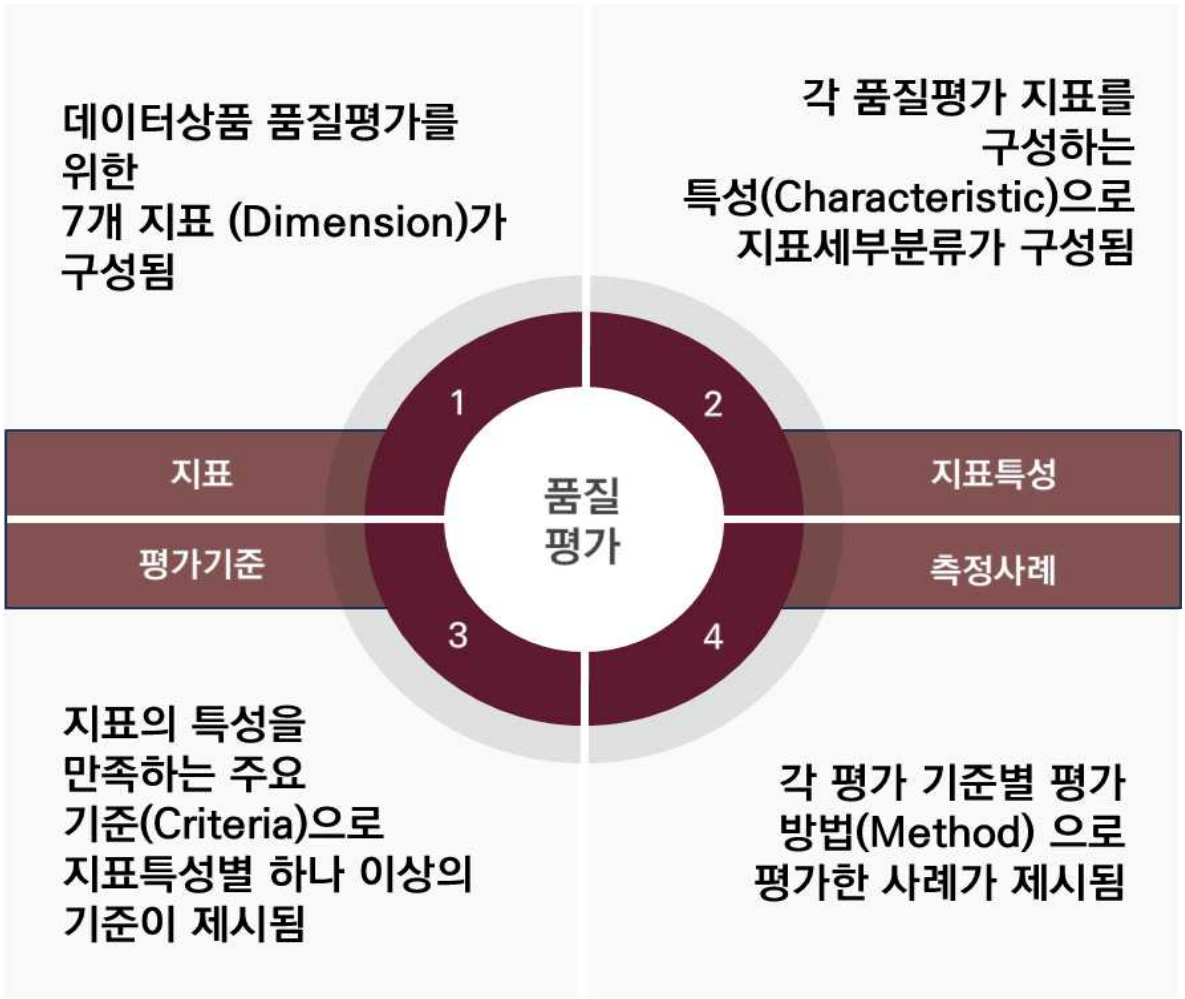
데이터상품의 품질평가지표는 객관성과 활용성을 확보하기 위하여 국내외 데이터거래소의 데이터상품 명세서와 데이터상품 샘플을 분석하였으며 데이터상품의 특성을 반영하기 위해 정형데이터와 비정형데이터로 분류하여 지표를 정의하였다.
데이터상품의 품질평가지표는 총 7개의 지표대분류와 22개 정형세부지표, 13개 비정형 세부지표 특성으로 구성되어 있다.

- 데이터상품의 유형인 정형, 비정형 데이터의 품질을 정량적으로 평가하기 위한 기준
- 데이터상품을 관리하기 위한 관리 메타데이터의 평가 기준
- 데이터상품 항목, 레코드, 파일이 충분히 제공되고 있는지에 대한 기준
- 데이터상품의 서비스가 충분히 제공되고 있는지에 대한 기준

(2) 구성

품질평가지표는 7개 지표(완전성, 유효성, 일관성, 정확성, 유일성, 적시성, 활용성)와 29개

지표특성(정형/비정형 동일지표 포함)으로 분류되고 각각의 지표특성에 대해 측정기준과 측정사례 등으로 구성된다.



[그림 2-8] 품질평가지표분류

(3) 정의

- **완전성(Completeness) 지표**
  - 완전성 지표는 데이터상품의 생성하는데 논리적인 상품 설계도인 데이터상품명세서와 물리적인 데이터상품이 누락 없이 설계·생성되었는지를 평가하는 지표이다.

[그림 2-9] 데이터상품 품질평가지표

- 이 지표에는 데이터상품 항목(Attribute), 레코드(Record), 파일(File), 물리메타데이터(Physical Meta Data), 관리메타데이터(Management Meta Data)의 특성으로 세분화하여 평가지표가 제시되어 있다.

- **유효성(Validity) 지표**
  - 유효성 지표는 데이터상품이 정의된 기준에 맞게 유효한 정보의 범위와 형식으로 되어 있는 수준, 데이터의 기능이 유효하게 서비스되는지에 대한 관리 수준 등을 평가하는 지표이다.
  - 이 지표에는 범위(Range), 형식(Format), 목록(List), 응답(Response), 데이터 기능(Data Function) 의 특성으로 세분화하여 평가지표가 제시되어 있다.
- **정확성(Accuracy) 지표**
  - 정확성 지표는 데이터상품이 실제 메타데이터에서 정의한 대로 정확하게 입력되었는지, 실제 입력된 값이 업무적 요건에 맞게 저장되어 있는지 등을 측정하는 지표이다.
  - 이 지표에는 메타데이터(Meta Data), 의미(Semantic), 계산/집계(Calculation/Aggregation), 선후관계(Order Relationship), 파일오류(File Error), 내용오류(Content Error)의 특성으로 세분화하여 평가지표가 제시되어 있다.
- **일관성(Consistency) 지표**
  - 일관성 지표는 항목, 레코드, 파일 상호 참조관계에 대한 일관성 확보수준 또는 메

타데이터와 데이터상품 파일 간 일관성이 유지되고 있는지를 측정하는 지표이다.

- 이 지표에는 참조무결성(Reference integrity), 항목 형식(Attribute Format), 항목 값(Attribute Value), 항목 관계(Attribute Relation), 메타데이터(MetaData)의 특성으로 세분화하여 평가지표가 제시되어 있다.

- **유일성(Uniqueness) 지표**
  - 유일성 지표는 항목, 레코드, 파일 중복으로 인한 모순의 위험성 확률을 평가하는 지표이다.
  - 이 지표에는 데이터상품 항목(Attribute), 레코드(Record), 파일(File)의 특성으로 세분화하여 평가지표가 제시되어 있다.
- **적시성(Timeliness) 지표**
  - 적시성 지표는 사용자가 만족하는 수준의 응답시간으로 데이터상품이 제공되었는지, 데이터상품 요청으로부터 수집·처리되어 제공되기까지의 작업시간이 최소화 관리되며 요구된 정보가 최신의 것인가에 대한 수준을 측정하는 지표이다.
  - 이 지표에는 응답시간(Response Time), 데이터 제공(Lead time), 최신값(Current Value)의 특성으로 세분화하여 평가지표가 제시되어 있다.
- **활용성(Relevance) 지표**
  - 활용성 지표는 사용자가 만족하는 수준의 충분한 정보가 제공되고 있는가, 데이터상품에 접근이 사용자의 편의성이 확보되었는가, 사용자가 정보를 유용하게 활용하고 있는가에 대한 수준을 측정하는 지표이다.
  - 이 지표에는 친밀(Intimacy), 효율(Efficiency), 활용(Usability)의 특성으로 세분화하여 평가지표가 제시되어 있다.

3.6. 품질평가 방법

효율적인 데이터상품 품질평가를 위한 평가 방법으로는 프로파일링, 업무규칙, 기계학습, 비정형실측 등이 있다.
또한 평가대상에 대한 처리 방법으로는 1. 직접 평가, 2. 적재 평가 방식으로 구분될 수 있다.

[그림 2-10] 데이터상품 품질평가 방법 분류

[표 2-5] 데이터상품 품질평가 방법 세부

품질평가 방법		내 용
프로파일링	항목 단위평가	○ 데이터 값의 완전성, 유효성, 정확성 등 데이터 항목 자체 오류를 분석하는 방법 ○ 컬럼분석, 날짜분석, 패턴분석, 코드분석 등을 통해 데이터 값의 정확성을 중심으로 평가
	레코드 단위평가	○ 데이터 값의 일관성, 유일성 등을 확보하지 못하는 결함을 분석하고 평가하는 방법 ○ 항목, 레코드, 파일간의 관계 정의 등 데이터의 구조적 결함 측정
	메타데이터 평가	○ 데이터상품 명세서의 관리메타데이터와 객체메타데이터의 완전성, 유효성, 정확성 등을 분석하여 평가하는 방법
업무규칙		○ 데이터상품 생성 시 정의된 업무기준(산출식)에 근거하여 데이터가 관리되고 있는지를 평가하는 방법 ○ 업무규칙(BR; Business Rule)을 준수하고 있는지에 관한 측정 스크립트(정규표현식, 프로그램, Function, SQL등)를 실행하여 오류 값을 추출
기계 학습		○ 빅 데이터 분석 방법 중 하나로 대량의 데이터를 이용하여 사람이 인지하는 패턴을 학습하고 이를 활용하여 사람의 개입 없이 기계(컴퓨터)에 의해

	데이터를 사람과 같이 판단 또는 예측 하는 방법
비정형 실측 (육안 평가)	○ 문서, 이미지, 동영상 등 정형화되어 있지 않는 정보를 사람이 직접 확인(실측)을 통하여 오류 여부를 평가하는 방법 ○ 별도 도구 없이 직접 정보를 조회하거나 해당 문서를 수기로 확인 등

[표 2-6] 평가대상 접근방식에 따른 평가방식

품질평가 방법	내 용
직접 평가	○ 데이터상품 파일 또는 API를 직접 읽어 프로그램을 이용하여 평가하는 방식
적재 평가	○ 데이터상품 파일 또는 API를 상품명세서의 항목정의서를 기준으로 적재시스템의 DB에 적재 후 SQL 및 프로그램을 이용하여 평가하는 방식

(1) 프로파일링 진단

- 프로파일링 평가는 생성, 제공, 활용중인 데이터상품 파일을 대상으로 데이터상품명세서 및 데이터상품 파일이 표준형식의 기준에 따라 설계·저장·제공되어 있는지를 분석하여 오류를 추정하는 진단 방법이다.
  - 데이터 프로파일링 방법으로 주로 프로파일링 도구를 활용하여 관리메타 분석, 통계분석, 분포분석, 이상치분석, 빈도분석, 날짜분석, 패턴분석, 코드분석, 구조 관계 분석 등을 진단한다.

(2) 업무규칙 진단

- 업무규칙 진단은 데이터상품의 업무적인 기준(산출식)에 맞게 실제 데이터상품 파일에 데이터 값이 저장되어 있는지를 진단하는 방법이다.
  - 업무규칙(BR)을 준수하고 있는지에 대한 평가 규칙을 스크립트(정규표현식, SQL, Function, Application등)를 실행하여 오류 값을 추출하는데 적용한다.

(3) 기계학습

- 빅 데이터 분석 방법 중 하나로 대량의 데이터를 이용하여 사람이 인지하는 패턴을 학습하고 이를 활용하여 사람의 개입 없이 기계(컴퓨터)에 의해 데이터를 사람과 같이 판단 또는 예측하는 방법

(4) 비정형실측

- 비정형실측은 문서, 이미지, 동영상 등 정형화되어 있지 않은 정보를 사람이 직접 확인(실측)함으로써 오류 여부를 진단 방법이다.
  - 별도의 도구 없이 사용자가 직접 정보를 조회하거나 문서 확인을 하는 방법

3.7. 품질평가 절차

품질평가절차는 데이터상품의 품질 수준을 파악하고 품질 제고를 위한 진단과 개선 방안을 체계화한 것으로 평가준비, 품질평가, 평가분석, 품질개선 등 4개의 구성 절차와 8개의 활동으로 구성되어 있다.

[그림 2-11] 데이터상품 품질평가 절차

3.8. 품질평가 준비

(1) 환경 분석

데이터상품의 품질평가를 위한 환경 분석(평가 요구사항, 평가대상 데이터상품 파일의 환경, 평가 수행여건과 규모 등)을 통해 적용 가능한 평가 환경을 식별한다.

가) 비즈니스 환경 분석

- 품질평가 담당자는 데이터상품 생성·제공·활용단계에 따른 품질이슈와 요구사항을 조사·분석하여 평가대상 데이터상품을 정한다.
  - 시간, 비용, 인력 등 자원의 제약을 감안하여 실제 수행 가능한 데이터상품 범위를 선정함

나) 평가 환경 분석

- 평가대상 데이터상품 파일의 구성 및 현황, 데이터 입력 및 활용 현황, 데이터상품 서비스 현황 등 품질 이슈와 관련된 데이터상품의 운영 환경을 파악한다.

- 데이터상품의 제공방식(파일, API, 서비스)에 따라 데이터상품의 직접평가방식, 수집 및 적재를 통한 평가방식 등의 방식을 선정함

다) 품질평가 수행계획 작성

- 품질평가 실현을 위한 구체적인 수행계획을 수립한다.
  - 품질평가 목적 및 목표 설정
  - 품질평가 범위 선정
  - 품질평가 조직의 책임과 역할 정의
  - 품질평가 시스템 분석
  - 품질평가 H/W, S/W 자원 파악
  - 품질평가 소요예산 및 수행 기간 정의

(2) 데이터상품 명세서 수집 및 정의

가) 상품명세서 정의

- 품질평가를 위한 데이터상품의 명세서를 수집하거나 작성한다.
  - 기본정보의 필수 항목 입력
  - 컬럼정의 항목 입력
  - 코드정의 코드명세 입력

구분	내용	비고
상품카테고리	IT/과학기술	필수
상품명	LTE 유동인구 데이터	필수
상품설명	□ 개요 o KT LTE 시그널 데이터 기반으로 50셀 단위로 유동인구 데이터 집계 □ 특징 o 월단위 제공 o 원천데이터 - KT □ 상품 제공 범위 o 추가 데이터 문의 가능	필수
과금유형	무료	필수

구분	내용	비고
상품가격	0원	필수
가격정책	이 상품은 무료 상품입니다.	
사용언어	한국어	
상품형태대분류	정형	필수
상품형태소분류	문서	필수
확장자	CSV	필수
압축형식	해당없음	
패키지형식	해당없음	
제공방식	Download	필수
사이즈	926KB	
건수	3996건	필수

[그림 2-12] 기본정보 정의 예시

(3) 데이터상품 유형 및 평가지표 식별

가) 데이터상품 유형

- 환경 분석 결과를 통해 평가대상 데이터상품 유형과 적용 가능한 평가지표를 식별하고 품질평가 수행 여건과 규모를 고려하여 품질평가지표별 기능과 방법을 검토한다.

나) 데이터상품 유형 식별

- 데이터상품을 형태적 분류(정형, 비정형)와 데이터상품의 특성인 데이터의 제공방식(파일, API, 다운로드, 서비스)을 포함하여 데이터상품의 유형을 분류하고 각 유형의 특성에 따른 중점관리 지표를 파악한다.
  - 데이터상품유형과 그에 따른 중점관리 지표는 품질평가결과에 대한 품질실태 분석 관점으로 활용됨

다) 데이터상품 항목별 도메인 정의

- 정형 데이터일 경우 상품명세서의 컬럼정의서 항목별로 도메인 그룹과 세부 도메인을 정의한다.

라) 데이터상품 항목별 유효기준 및 평가지표 정의

- 정형 데이터일 경우 상품명세서의 컬럼정의서 항목별로 유효기준을 정의한다.

[표 2-7] 유효/오류 기준 정의

컬럼한글명	데이터타입	길이	도메인	유효/오류 기준
셀ID	VARCHAR	8	명	ZZZ-CC-CC-ZZZ
기준년월	VARCHAR	6	일자	YYYYMM
X좌표	NUMBER	18,9	수	속성 > -99 AND 속성 < 120
Y좌표	NUMBER	18,9	수	속성 > -99 AND 속성 < 120
요일	VARCHAR	1	코드	0,1,2,3,4,5,6
시간대	VARCHAR	2	코드	0,1,2
남자05	NUMBER	10	수	속성 > 0 AND 속성 < 100000
남자10	NUMBER	10	수	속성 > 0 AND 속성 < 100000
남자15	NUMBER	10	수	속성 > 0 AND 속성 < 100000
여자05	NUMBER	10	수	속성 > 0 AND 속성 < 100000
여자10	NUMBER	10	수	속성 > 0 AND 속성 < 100000
여자15	NUMBER	10	수	속성 > 0 AND 속성 < 100000
합계	NUMBER	10	수	속성 > 0 AND 속성 < 100000
행정동코드	VARCHAR	8	코드	행정동.행정동코드
행정동이름	VARCHAR	100	명	[ㄱ-ㅎㅏ-ㅣ]+$

3.9. 품질평가 실시

(1) 지표별 품질평가

- 품질평가 기관 또는 기업이 보유한 데이터상품의 품질 수준을 직접 측정하는 단계로, 측정도구별 절차 수행을 통해 데이터상품의 오류(추정)를 도출한다.
- 데이터상품의 항목별 유효기준 및 평가기준에 따른 점검 방법을 정의한다.
- 데이터품질 점검 방법은 정형데이터의 경우 직접 평가와 적재 평가 방식에 따라 평가대상 데이터를 준비한다.

【적재 평가절차】

데이터베이스 적재
- 데이터상품 정의서의 컬럼정의에 따라 테이블을 생성
- CSV, EXCEL과 같은 파일일 경우 로더를 통해 데이터를 적재
- API일 경우 JSon데이터를 CSV로 변환하여 데이터를 적재
- 데이터상품 정의서를 적재

평가대상 테이블 선정
- 품질평가 대상 상품 중 실제 측정할 대상 상품을 적재한 테이블을 선별함

항목별 평가방법 정의
- 데이터상품 항목별 유효기준 및 평가지표 정의 내용에 따라 항목별로 품질평가 방법(SQL 작성 등)을 상세히 정의

품질평가 실행
- 품질평가 실행 일정에 따라 평가를 수행함
- 품질평가 실행은 품질평가 솔루션, 프로그램, SQL등을 활용하여 실행할 수 있으며 평가 결과는 어떤 도구를 사용하든 동일해야 함

품질평가 결과보고서
- 품질평가 수행 결과 보고서를 작성함
- 품질평가 결과보고서는 데이터상품별, 속성별, 도메인별 오류데이터의 집계 및 상세 데이터를 포함하여 작성함

(2) 평가 결과 지수화

- [표2-1] 품질지수 산정기준에 따라 전체 건수 대비 오류건수에 대한 오류율(%), 정합성율(%)을 산정하고, 6시그마 관점에서 DPMO(EPM), 6시그마지수, 품질지수를 산정한다.
  - 지표별 오류율(%) = (오류건수 ÷ 진단건수) × 100
  - 종합 오류율(%) = (오류전체건수 ÷ 진단전체건수) × 100

3.10. 품질평가 결과분석

(1) 오류유형 분류

- 평가지표별 평가결과로부터 오류 내역을 파악하여 오류 유형별 오류발생 원인을 분석한다.

(2) 데이터상품 원천 조사

- 오류의 원인분석 및 개선 수행을 위해 상품의 원천데이터를 조사한다.
- 데이터 오류가 데이터상품 생성 시 다양한 문제로 인해 원천데이터와는 다르게 생성될 수 있으므로 실제 원천데이터와 상품데이터의 실데이터가 일치하는지 조사한다.

3.11. 품질평가 오류개선

(1) 개선(안) 도출

- 데이터 오류가 명확히 규명된 오류원인과 그에 따른 서비스 파급효과 및 데이터상품 원천의 데이터 품질관리 이슈에 대한 종합적인 시사점을 도출한다.
  - 중점관리 지표 또는 상대적으로 오류가 많은 지표 측면
  - 평가지표별 오류발생 원인 측면
  - 업무영향 범위(데이터상품의 현재 서비스 상태) 측면
  - 원천데이터의 품질관리 수준 현황 측면
  - 품질관리 비용적/업무적 시급과제 측면

(2) 품질개선 계획 수립

- 품질평가 결과에 따라 개선 권고안이 도출되면 담당자는 실제 개선 수행을 위한 품질개선 계획을 수립해야 한다.
- 개선과제 도출
  - 품질평가 결과 오류유형에 따라 개선방안을 구체화하여 개선과제를 도출한다.

- 개선과제 우선순위 부여
  - 개선과제는 데이터상품의 서비스 시급성과 중요성 등을 고려하여 장기, 단기 개선과제로 분류하여 수행 우선순위를 부여한다.

(3) 데이터상품 개선

- 개선(안)과 품질개선 계획에 따라 담당자는 데이터상품의 품질을 개선한다.
- 필요할 경우 데이터상품의 원천데이터에 대한 개선을 실시한다.

PART 04

Korea Data Agency

정형데이터

4.1. 개요

데이터상품은 개방데이터나 내부 업무데이터와 달리, 특정 다수 고객에 대한 유료 서비스를 목표로 하므로 데이터품질이 매우 중요하다.
공급자와 수요자에 의한 거래이고 플랫폼에 의해 거래가 이루어지기 때문에 평가 모형이 단순하고, 정량적으로 측정 가능해야 하며, 쉽게 시스템화할 수 있어야 한다. 본 가이드에서는 데이터품질 전문가의 도움 없이 데이터공급자와 거래소, 수요자에 의해 쉽게 정량적으로 평가하고 자동화하는 것을 목표하고 있다.

4.2. 정의

데이터상품 품질평가의 주체, 방법, 평가지표 등에 대하여는 생성주체별, 거래소 또는 담당자별로 접근방법이 상이할 수 있다.
품질평가 방법은 거래소의 품질관리 수준을 고려해서 결정되어야 하므로, 일반적으로 통용될 수 있는 방법은 없다.
그러나 데이터상품 품질평가를 성공적으로 실행하기 위해서는 우수한 품질평가 인력, 적정한 품질평가 지표, 국가 차원의 법령 등의 요건이 갖추어져야 한다.
품질평가자는 조직 내·외부에서 품질평가의 각 부문별로 에러 발생요인 및 평가에 대한 전문지식을 갖춘 사람으로 선정해야 한다.
품질평가지표는 평가대상 데이터상품별로 달라져야 하며, 명료성, 지속적 측정가능성, 평가 기술적 용이성등의 요건을 갖춘 지표가 바람직하다.

4.3. 대상

본 절에서는 구조화, 반구조화되어 있는 정형데이터상품에 대한 품질평가를 대상으로 하며 반드시 상품명세서를 작성하는 것을 기준으로 한다.

4.4. 품질평가지표 구성

앞서 정의된 7가지의 품질평가지표 대분류를 기준으로 정형 데이터 관련 22가지의 세부 지표를 아래 [표2-8]과 같이 정의한다. 세부 품질평가지표들은 국내·외 선진사례 및 논문 13종의 현황분석을 통하여 시사점을 도출하고 데이터상품의 특성을 고려하여 다음의 지표를 도출하였다.

[표 2-8] 품질 진단 세부 지표

대 분류	세부 분류	정의
완전성	항목 완전성	○ 필수 항목은 반드시 값이 입력되어야 한다.
	레코드 완전성	○ 반드시 있어야 하는 레코드는 반드시 존재해야 한다.
	상품 완전성	○ 상품 파일, API, URL 은 반드시 존재해야 한다.
	메타 완전성	○ 메타데이터의 필수 값은 반드시 입력되어야 한다.
유효성	범위 유효성	○ 데이터 값은 유효 범위 내 값이어야 한다.
	형식 유효성	○ 데이터 값은 유효 형식의 포맷을 준수해야 한다.
	목록 유효성	○ 데이터 값은 유효 값의 목록에 존재하는 값이어야 한다.
	응답 유효성	○ API, 다운로드, 서비스상품 호출 시 응답이 되어야 한다.
정확성	메타 정확성	○ 메타데이터에 기재된 값과 실제 등록된 값은 일치해야 한다.
	의미 정확성	○ 데이터 값은 업무적으로 의미 있는 값이어야 한다.
	계산/집계 정확성	○ 항목 간 계산에 의해 산출되었을 경우 계산 결과는 정확해야 한다.
	선후관계 정확성	○ 선후관계를 갖고 있을 경우 선행과 후행이 정확해야 한다.
적시성	데이터 제공 적시성	○ 데이터는 갱신주기에 따라 정상적으로 갱신되어야 한다.
	데이터 최신성	○ 데이터는 최근 데이터를 포함하고 있어야 한다.
	응답 적시성	○ API/서비스 호출 시 송수신 속도가 활용 가능한 수준이어야 한다.
일관성	참조 무결성	○ 하나 이상의 상품이 상속관계를 갖고 있을 경우 참조 상품은 기준상품에 반드시 정의된 값이어야 한다.
	항목 형식일관성	○ 하나 이상의 상품의 동일한 데이터 항목은 동일한 데이터 형식으로 관리되어야 한다.
	항목 값 일관성	○ 하나 이상의 상품의 동일한 항목은 동일한 데이터 값으로 관리되어야 한다.
	항목 관계 일관성	○ 항목간에 연관관계가 있는 경우 서로 일관되게 관리되어야 한다.
유일성	항목 유일성	○ 식별자 단일 항목은 유일해야 한다.
	레코드 유일성	○ 식별자 항목을 기준으로 레코드는 유일해야 한다.
활용성	개인정보 익명성	○ 개인정보 항목은 비식별화되어야 한다.

4.5. 품질평가지표 상세

(1) 완전성 (세부지표 : 항목, 레코드, 상품, 메타)

데이터상품의 항목, 레코드, 파일, 메타데이터의 필수 항목은 누락 없이 입력되어 있어야 하며 충실하게 관리 되어야 한다.

가) 항목 완전성

① 개요

[표 2-9] 항목 완전성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 상품의 데이터 항목 중 반드시 값이 입력되어 있어야 하는 항목의 누락 없이 잘 관리되는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 정형 데이터상품 파일의 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 단독 완전성 : 데이터상품 파일의 항목 또는 API호출 결과 json array의 데이터셋 Key의 Value가 필수일 경우 값이 반드시 존재해야 함
(예시) LTE유동인구.csv 데이터상품 파일의 “시간대” 항목은 반드시 존재해야 함
○ 조건 완전성 : 다른 항목의 값에 상태에 따라 반드시 존재해야 함
(예시) 결혼여부가 "Y"이면 결혼기념일은 반드시 존재해야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 항목의 데이터 값이 NULL 이거나 공백인 건수
○ 분모기준(B) : 항목의 전체건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [지역별 카드매출현황 상품]의 "거래시도" 항목 값이 반드시 존재해야 하나 공백이 존재함
○ [지역별 카드매출현황 상품]의 "고객구분" 항목 값이 "개인" 일 경우 "연령대"는 반드시 존재해야 하나 공백이 존재함</td></tr>
</table>

② 필요성

정형 데이터상품의 구성요소는 가장 작은 단위 항목(속성)과 항목(속성)의 집합인 레코드, 레코드의 집합인 파일, 파일의 집합인 상품과 데이터상품의 구성요소에 대한 명세인 상품명세서로 구성된다. 이러한 구성요소의 필수 항목들은 누락 없이 생성되고 제공되어야 한다. 항목 완전성은 데이터상품의 가장 작은 단위인 필수 항목의 누락을 방지하고 충실도를 높이기 위함이다.

③ 평가 방법

- 평가대상 항목이 NULL인 경우 오류로 평가
- 평가대상 항목이 공백일 경우 오류로 평가할지 여부 판단

나) 레코드 완전성

① 개요

[표 2-10] 레코드 완전성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 상품의 레코드(ROW)가 누락 없이 잘 관리되는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 특정한 조건에 따라 필수적으로 생성되어야 하는 데이터상품
1. 시간 주기데이터 : 특정 날짜 주기별 생성되는 레코드 데이터상품
2. IOT 데이터 : 특정 이벤트 또는 주기적인 데이터 발생에 의해 생성되는 데이터상품
3. 조건 데이터 : 지역, 분류등 특정 디멘전이 존재하여 발생하는 데이터상품</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 시간 주기데이터 : 년,월,일,시,분,초별과 같이 생성주기가 있는 데이터상품의 경우 생성주기에 누락이 발생할 수 없음
(예시) 일별 데이터 발생 추이 데이터에 특정일 데이터가 누락될 수 없음
○ IOT : 센서데이터, 로그데이터 주기적 수집 시 해당 주기/영역에 누락이 발생할 수 없음
(예시) 17개 시도 미세먼지 수치에 17개 시도 레코드가 모두 존재해야 함</td></tr>
</table>

품질지수 계산	○ 분자기준(A) : 식별자, 디멘전, 발생주기등 반드시 존재해야 하는 레코드가 누락되어 있는 레코드 건수 ○ • 분모기준(B) : 레코드의 전체건수 ○ • 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
오류 예시	○ [2020년 8월 일자별 카드매출현황 상품]에 "매출일자"가 20200821인 매출현황 레코드의 누락이 존재함 ○ 센서데이터에서 수집주기의 레코드 누락이 존재함

② 필요성

레코드 완전성은 데이터항목이 모여 정보로서 가치를 나타내는 단위로 데이터의 발생 주기, 조건등에 따라 데이터가 반드시 입력되어 있어야 한다. 그러나 특정 날짜의 집계데이터 누락, 특정 센서의 전송데이터 누락등의 레코드 누락으로 인하여 시계열 데이터 분석 시 품질에 문제를 발생시킬 수 있으므로 반드시 검사해야 하는 항목이다.

③ 평가 방법

- 레코드가 기준일자(YYYYMM 포맷)를 기준으로 집계 데이터 제공 시 기준일자 누락이 존재할 수 없음을 평가

다) 파일 완전성

① 개요

[표 2-11] 파일 완전성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">상품의 파일 또는 API가 누락 없이 등록되는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 파일상품 : 데이터상품의 파일
○ API상품 : API URL</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형
비정형</td><td>파일
API</td><td>제공자
거래소
구매자</td><td>정량
(시스템)</td><td>운영</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 파일상품 : 데이터상품이 파일일 경우 파일이 반드시 존재해야 함
(예시) 파일이 등록되지 않음 / 파일의 정상다운로드 여부는 “응답 유효성” 에서 점검함
○ API상품 : 데이터상품이 API 상품일 경우 API 규약과 API URL이 모두 존재해야 함
(예시) API 규약이 등록되지 않음 / API의 정상호출 여부는 “응답 유효성” 에서 점검함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 상품명세서에 등록된 파일/API URL의 전체 개수 - 등록된 파일/API URL의 전체 개수
○ 분모기준(B) : 상품명세서에 등록된 파일/API URL의 전체 개수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
○ 상품명세서에 파일개수, API개수가 누락되어 있을 경우 현재 등록된 파일, API개수를 분모/분자 기준으로 정의</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [지역별카드매출현황 상품]의 CSV 파일이 등록되지 않음
○ [미세먼지 수치 API 상품]의 URL이 누락되어 있음</td></tr>
</table>

② 필요성

정형 데이터상품은 제공방식에 따라 파일, API, 서비스등을 통해 구매자에게 제공되고 활용된다. 이러한 파일 또는 API, 서비스가 상품명세서에 정의된 내역과 다르게 누락이 발생하고 있을 경우 상품서비스에 심각한 장애를 초래할 수 있다. 따라서 공급자는 상품 등록 시

반드시 상품명세서를 기준으로 파일의 누락을 검증해야 하며 거래소는 운영 중인 데이터 상품의 파일 누락을 주기적으로 점검해야 한다.

③ 평가 방법

- 상품명세서의 파일건수와 실제 등록된 파일 개수가 동일한지 평가
- 평가를 위해 상품명세서와 상품파일을 DBMS에 테이블로 생성하여 적재 후 평가 실시
- 상품명세서는 {상품명세서} 테이블에 저장 / 상품파일은 {대상상품_파일명}에 저장

라) 메타 완전성

① 개요

[표 2-12] 메타 완전성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 상품 등록 시 등록되어야 할 메타데이터의 필수 값이 누락 없이 잘 관리되는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 상품명세서 : 데이터상품의 설명서인 명세서 파일
○ 상품명세서 : 상품명세서에 반드시 채워져 있어야 하는 메타정보
○ 항목정의서 : 항목정의서에 반드시 채워져 있어야 하는 메타정보</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형
비정형</td><td>파일
API</td><td>제공자
거래소
구매자</td><td>정량
(시스템)</td><td>운영</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 상품명세서 : 상품명세서에 필수항목은 반드시 입력되어야 함
(상품명세서 정의서 참조필요)
○ 항목정의서 : 항목정의서에 필수항목은 반드시 입력되어야 함
(상항목정의서 참조필요)
○ 상품명세서 : 데이터상품은 반드시 상품명세서가 반드시 존재해야 함</td></tr>
<tr><td>품질지수
계산</td><td colspan="7">○ 분자기준(A) : 상품명세서/항목정의서 필수 항목 누락 수
○ 분모기준(B) : 전체 카탈로그/항목정의서 필수 항목 수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
○ 상품명세서 파일이 누락되어 있을 경우 0점</td></tr>
<tr><td>오류
예시</td><td colspan="7">○ [지역별카드매출현황 상품]의 상품명세서 정보 중 "파일종류" 항목이 누락되어 있음
○ [지역별카드매출현황 상품]의 항목정의서 중 "도메인" 항목이 누락되어 있음</td></tr>
</table>

② 필요성

데이터상품을 공급하는 공급자, 운영하는 거래소, 활용하여 서비스를 제공하는 구매자는 상품명세서를 통해 데이터의 유형과 내용, 구조등을 파악할 수 있다. 따라서 상품명세서 파일은 반드시 존재해야 하며, 명세서의 필수 항목으로 지정된 항목에는 누락 없이 내용이 기술되어 있어야 한다.
구매자가 데이터상품을 검색 또는 구매하고자 할 경우 다양한 상품 정보 제공을 통해 정확한 상품 파악이 가능하도록 지원하기 위함이다.

③ 평가 방법

- 상품명세서의 필수 항목에 누락이 없는지 점검
- 상품명세서는 {상품명세서} 테이블에 저장

(2) 유효성 (세부지표 : 범위, 형식, 목록, 응답)

데이터상품의 항목은 데이터상품 활용에 필요한 유효한 범위, 형식, 목록의 값으로 저장되어야 하며 정해진 규약대로 데이터상품이 전송, 다운로드, 서비스 화면으로 유효하게 제공되어야 한다.

가) 범위 유효성

① 개요

[표 2-13] 범위 유효성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ API, 다운로드, 서비스상품 호출 시 정해진 규약대로 데이터상품이 전송되거나 다운로드 되거나, 서비스 화면으로 정상 이동되는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ API : 호출 URL
○ 다운로드 : 다운로드 URL
○ 서비스 : 서비스 URL</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
</table>

평가 기준	○ 날짜범위 - 날짜데이터의 FROM~TO 조건에 따라 크고 작은 날짜의 순서가 유효해야 함 ○ 숫자범위 - 숫자데이터의 FROM~TO 조건 / 최소값~최대값에 따라 크고 작은 숫자의 순서가 유효해야 함 ○ Outlier - 숫자데이터의 특이치 검사를 통해 Lo fense ~ Hi fense 범위를 벗어나는 특이치를 관리해야 함 대부분 정규분포에서 97.5% 이상 또는 2.5%의 이하에 포함되는 값을 이상치로 판별 *[범례] Q1 : 1사분위 , Q3 : 3사분위 , IQR : Q3 - Q1 , AVG : 평균 , VAR : 표준편차* *Test for Outliers (Median and Interquartile Deviation Method) : Lo fense (Q1－1.5 × IQR) / Hi fense (Q3 + 1.5 × IQR)* *Test for Outliers (Mean and Standard Deviation Method) : Lo fense (AVG－1.5 × VAR) / Hi fense (AVG + 1.5 × VAR)*
품질지수 계산	○ 분자기준(A) : 유효범위 위배 건수 ○ 분모기준(B) : 항목의 전체건수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
오류 예시	○ [2020년 8월 일자별 카드매출현황 상품]에 "매출일자"가 2020년 7월 29인 데이터가 존재함 ○ 지역별카드매출현황 상품]의 "거래금액"에 0보다 작은 금액이 존재함 (이상치 점검을 포함)

② 필요성

데이터상품의 항목 중 날짜, 수치(숫자) 데이터는 공급자의 업무적 기준 또는 범용적으로 정해진 유효한 범위를 갖고 있다. 범위 유효성 평가는 유효범위를 벗어난 값이나 이상치 데이터를 진단하고 이를 통해 수치(숫자) 데이터의 통계, 집계 오류등을 최소화하기 위함이다.

③ 평가 방법

- 상품의 날짜, 수치(숫자) 데이터가 유효한 범위 내에 있는지 평가
- 상품파일은 {대상상품_파일명}에 저장
- 범위 검증을 위해서는 항목의 최소값과 최대값의 유효범위가 지정되어 있어야 함

나) 형식 유효성

① 개요

[표 2-14] 형식 유효성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터 값이 가질 수 있는 유효 형식을 준수하고 있는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 날짜 항목 / 문자열 항목 / 번호 항목 / 금액, 수량, 율 항목
○ XML / JSON</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 날짜 항목 - 날짜포맷에 맞는 유효한 날짜가 입력되어야 함 (최초, 최종 날짜는 '99999999'와 같이 표준이 존재할 수 있음)
○ 문자열 항목 - 유효한 문자열(한글, 영문, 숫자 등의 조합) 구성에 대해 유효 문자열로만 구성되어야 함
○ 번호 항목 - 유효한 번호 포맷에 대해 채번규칙 및 포맷이 유효해야 함
○ 금액,수량,율 항목 - 숫자로만 구성되어야 하며 단위가 일관되어야 함
○ XML - DTD 유효성 점검을 통과해야 함
○ JSON - KEY/VALUE의 구조를 지켜야 하며 "(쌍따옴표)로 데이터가 구분되어야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 유효형식 위배 건수
○ 분모기준(B) : 항목의 전체건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [2020년 8월 일자별 카드매출현황 상품]에 "매출일자" 항목에 “2020899” 날짜포맷 오류데이터가 존재함
○ [2020년 8월 일자별 카드매출현황 상품]에 "거래지역"에 비완성형 한글이 존재함
○ [2020년 8월 일자별 카드매출현황 상품]에 "거래지역번호"는 5자리의 숫자로 구성되어 있어야 하나 4자리숫자가 존재함
○ [2020년 8월 일자별 카드매출현황 상품]에 "매출금액"은 원화로 기재되어 있어야 하나 "외화"로 기재되어 있음</td></tr>
</table>

② 필요성

데이터상품의 파일은 RDBMS와 달리 데이터의 타입을 지정하여 유효성을 보장받지 못한다.

따라서 수치데이터를 제공함에 있어 NUMERIC타입이 아닌 다양한 문자 타입등이 입력될 수 있는 가능성이 많다. 따라서 형식 유효성은 수치데이터의 문자 포함여부, 번호, 날짜데이터의 생성규칙 및 패턴, 문자열 데이터의 깨진 문자열 점검등을 통해 데이터상품 서비스 시 발생할 수 있는 오류를 방지하기 위함이다.

③ 평가 방법

- 상품의 포맷이 유효한 형식인지 평가
- 상품 파일은 {대상상품_파일명}에 저장
- 포맷 검증을 위해서는 항목의 유효형식이 지정되어 있어야 함

다) 목록 유효성

① 개요

[표 2-15] 목록 유효성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터 값이 가질 수 있는 유효 목록을 준수하고 있는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 목록형 데이터값의 항목 - 분류, 코드 도메인</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 분류 : 예/아니오 등과 같이 2~3개의 정해진 유효한 값으로만 구성되어야 함
○ 코드 : 특정한 목록을 갖는 목록, 범위의 코드값 또는 명칭을 가져야 함
코드값을 갖고 있는 경우 코드정의서를 제공해야 하며, 코드정의서에 정의된 내용만 사용해야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 유효값 목록 위배 건수
○ 분모기준(B) : 항목의 전체건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [서울시 유동인구 데이터상품]의 "사용여부"는 "예/아니오"의 값이어야 하나 "Y" 값이 존재함</td></tr>
</table>

	○ [지역별 카드매출현황 상품]의 "고객구분"은 "개인/법인"의 값이어야 하나 "사업자" 값이 존재함 ○ [지역별 카드매출현황 상품]의 "고객상태구분코드"는 코드정의서에 "1-활동 / 2-중지 / 3-탈회" 로 정의하였으나 "01" 값이 존재함

② 필요성

데이터상품의 파일은 RDBMS와 달리 코드값이 아닌 명칭으로 제공하는 경우가 더 많이 발생하고 있다. (예: Y/N을 예/아니오 , M/F를 남자/여자)
목록이 정해져 있는 항목의 경우 유효한 목록값 이외의 데이터가 존재하는지 지속적으로 검증하여야 한다.

③ 평가 방법

- 상품의 코드성 데이터가 유효한 목록값으로 관리되고 있는지 평가
- 상품파일은 {대상상품_파일명}에 저장
- 목록 검증을 위해서는 항목의 유효한 값의 목록 또는 코드가 지정되어 있어야 함

라) 응답 유효성

① 개요

[표 2-16] 응답 유효성 정의

항 목	내 용						
지표 정의	○ API, 다운로드, 서비스상품 호출 시 정해진 규약대로 데이터상품이 전송되거나 다운로드 되거나, 서비스 화면으로 정상 이동되는지 평가함						
평가 대상	○ API : 호출 URL ○ 다운로드 : 다운로드 URL ○ 서비스 : 서비스 URL						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	정형 비정형	파일 API	거래소 구매자	정량 (시스템)	운영	-	-

		서비스					
평가 기준	○ API : 호출 URL이 응답코드 200으로 정상 응답되어야 함 ○ 다운로드 : 다운로드 URL 호출 시 파일이 정상적으로 다운로드 되어야 함 ○ 서비스 : 서비스 URL 호출 시 해당 서비스 URL로 정상 이동 되어야 함						
품질지수 계산	○ 분자기준(A) : 등록된 파일의 전체 파일/API URL의 비정상 응답 개수 (404 Not Found ..) ○ 분모기준(B) : 등록된 파일의 전체 파일/API URL 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수						
오류 예시	○ [미세먼지 수치 API 상품]의 API를 호출하였으나 리턴 코드가 정상이 아닌 오류코드를 전송함 (오류코드는 API 정의서에 기재됨) ○ [서울시 유동인구 데이터상품]의 다운로드 시 다운로드 오류가 발생함 ○ [월별 도소매가격정보]의 서비스 선택 시 "농산물유통정보" 서비스 화면을 정상적으로 호출하지 못함						

② 필요성

정형 데이터상품은 파일 또는 API 형태로 제공되며 파일을 경우 다운로드 방식이 주로 사용되고 있다. 이러한 파일 또는 API가 정상적으로 다운로드되지 않거나 API 호출에 문제가 있을 경우 심각한 문제를 초래할 수 있다. 따라서 응답 유효성은 파일의 다운로드, API 호출, 서비스 화면으로의 이동이 정상적으로 동작하는지 주기적으로 점검하고 조치해야 한다.

③ 평가 방법

- 각 상품의 제공방식에 따라 정상적으로 다운로드 또는 API 호출이 되는지 확인
- 정상 호출은 1 / 비정상 호출은 0 으로 판단

(3) 정확성 (세부지표 : 메타, 의미, 계산/집계, 업무규칙)

실세계에 존재하는 데이터상품의 원천 데이터와 동일한 데이터가 제공되어야 하며 업무적 요건에 맞는 데이터가 제공되어야 한다.

가) 메타 정확성

① 개요

[표 2-17] 메타 정확성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 메타데이터에 기재된 값과 실제 등록된 값이 정확한지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 1. 건수 / 2. 크기 / 3. 명칭 / 4. URL / 5. 항목개수 / 6. 항목의미 / 7. 항목타입/길이</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형
비정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 데이터건수 : 메타에 등록된 건수와 실제 파일/API 호출결과의 건수가 동일해야 함
○ 데이터크기 : 메타에 등록된 크기와 실제 파일 크기가 임계범위 내에서 동일해야 함
○ 항목개수 : 메타에 등록한 항목과 실제 항목 개수가 동일해야 함
○ 데이터 항목 타입 : 메타에 등록한 항목의 데이터타입(문자형, 숫자형, 날짜형)와 실제 항목의 데이터타입이 동일해야 함
○ 데이터 항목 길이 : 메타에 등록한 항목의 데이터 길이보다 실제 항목의 데이터 길이가 작아야 하며, 소수점이 일치해야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : ABS(메타데이터 작성건수 - 실제 데이터 건수)
○ 분모기준(B) : GREATEST(메타데이터 작성건수 , 실제 데이터 건수)
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [서울시 유동인구 데이터상품]의 상품명세서에 상품데이터 건수가 10,000건으로 등록되어 있으나 데이터는 9,435건이 존재함
○ [서울시 유동인구 데이터상품]의 상품명세서에 파일사이즈가 1Mbyte로 되어 있으나 30K의 크기로 존재함 (Threshold를 적용)
○ [서울시 유동인구 데이터상품]의 상품명세서에 파일명이 "2020년 8월 서울시 유동인구.csv"로 등록되어 있으나 "2020년 8월 서울시 유동인구_0.1.csv"로 파일명이 상이함
○ [미세먼지 수치 API 상품]의 상품명세서에 URL과 실제 호출 URL이 상이함</td></tr>
</table>

② 필요성

정형데이터의 상품명세서는 상품명세, 항목정의, 코드정의로 구성되며 공급자, 거래소, 구매자는 해당 내용을 통해 상품을 공급하고 활용한다. 따라서 상품명세서에 기술되어 있는 정보와 실제 데이터상품 파일의 내용이 불일치할 경우 구매자가 상품을 이용한 서비스 구현 시 다양한 문제들이 발생할 수 있다. 메타 정확성은 오류를 사전에 방지하기 위하여 상품명세서의 주요 항목과 실제 제공되는 파일, API가 정확히 일치하는지 주기적으로 점검해야 한다.

③ 평가 방법

- 상품명세서의 내용과 실제 데이터상품이 동일한지 평가
- 평가를 위해 상품명세서와 상품파일을 DBMS에 테이블로 생성하여 적재 후 평가 실시
- 상품명세서는 {상품명세서} 테이블에 저장 / 상품파일은 {대상상품_파일명}에 저장

나) 의미 정확성

① 개요

[표 2-18] 의미 정확성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터 값이 값으로서의 의미가 정확하지 않은지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 문자 입력 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>-</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 무의미한 문자열 : 데이터가 실제 사용될 수 있는 의미 있는 정확한 값이 입력되어야 함
(동일한 문자열로 단순 나열등으로 의미파악이 불가할 수 없음)</td></tr>
<tr><td>품질지수
계산</td><td colspan="7">○ 분자기준(A) : 무의미한 데이터 건수
○ 분모기준(B) : 항목의 전체건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류
예시</td><td colspan="7">○ [서울시 유동인구 데이터상품]의 "유동인구명"에 "11111", "가가가가가가가가"와 같이 의미없는 문자열이 존재함</td></tr>
</table>

② 필요성

사용자의 자유로운 입력데이터의 경우 무의미한 데이터를 입력하여 업무적으로 사용할 수 없는 경우가 다수 발생하고 있다. 따라서 명칭 또는 내용에 사용자가 무의미하게 입력한 데이터에 대해 검증을 실시해야 한다.

③ 평가 방법

- 내용 또는 명칭과 같은 사용자정의 입력데이터에 대해 동일 단어를 무작위로 입력하는 경우 검증

다) 계산/집계 정확성

① 개요

[표 2-19] 계산/집계 정확성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터 값이 항목간에 서로의 계산에 의해 산출되었을 때 정확한지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 금액/수량/율 : 상품 내 계산 및 집계가 존재하는 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 금액/수량/율 : 특정 항목간의 계산에 따라 집계나 산식 결과가 나오는 경우 해당 계산식의 결과가 정확해야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 계산/집계 불일치 건수
○ 분모기준(B) : 항목의 전체건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [지역별 소상공인 거래 상품]의 "상품금액" , "판매건수", "판매금액" 항목에서 "상품금액 " X "판매건수" = "판매금액"과 일치하는 값이 존재함</td></tr>
</table>

② 필요성

데이터상품에서 제공되는 수치데이터는 다양한 계산과 집계를 통해 제공되고 활용된다. 따라서 항목, 레코드, 파일간의 데이터 계산 및 집계 결과는 정확하게 제공되어야 한다.

③ 평가 방법

- 금액, 수량, 율등과 같은 수치데이터를 기반으로 특정한 계산식과 집계에 의해 발생되는 항목은 계산/집계 산식을 먼저 조사한다.
- 계산/집계값은 하나의 파일 또는 여러 파일의 계산/집계에 의해 결정될 수 있다.

라) 업무규칙 정확성

① 개요

[표 2-20] 업무규칙 정확성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터상품을 생성함에 있어 관리하는 데이터의 생성 규칙(산출식)으로 여러 데이터간의 관계에 의해 의미적인 데이터의 정확성을 진단하는 로직 임</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 데이터상품의 생성 근거가 되는 가이드나 지침·규정에 정의된 업무적인 기준(산출식) 대상 항목, 레코드, 파일</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 데이터의 생성 규칙을 정확히 준수해야 한다.</td></tr>
<tr><td>품질지수
계산</td><td colspan="7">○ 분자기준(A) : 업무규칙 위배 건수
○ 분모기준(B) : 전체건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류
예시</td><td colspan="7">○ 매일 정시에 기온은 측정되어야 한다. 데이터상품 API에 제공된 측정 시간이 매 정시인지를 확인하는 것은 업무 규칙을 확인하는 것임
○ 총급여는 총근로시간*시간당 최저임금 이상이어야 함</td></tr>
</table>

② 필요성

데이터상품은 단순한 데이터의 생성 규칙에 따라 생성되는 것이 아니라 업무적인 요건과 규정에 따라 생성되고 활용된다. 따라서 IT 관점의 데이터 오류 검출뿐만 아니라 Business 관점의 오류 검출이 병행되어야 한다.

③ 평가 방법

- 업무규칙은 생성자가 데이터상품의 특성에 따라 업무적인 절차나 규정의 준수 여부를 SQL 또는 프로그램 형태로 자유롭게 기술하여 평가

(4) 일관성 (세부지표 : 참조무결성, 항목형식, 항목값, 항목관계)

데이터상품의 동일한 항목의 형식, 값, 관계 및 참조관계가 일관되게 제공되어야 하며 데이터상품 간 표준, 항목 정의, 데이터 형식 등이 일치하여야 한다.

가) 참조무결성

① 개요

[표 2-21] 참조무결성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터상품의 한 항목이 다른 상품의 특정항목을 참조하고 있으면, 참조하는 상품에 동일한 값을 포함하는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 상품 : 데이터상품의 파일이 2개 이상 존재해야 하며, 데이터상품의 파일간 관계정보가 필요</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>식별자</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 상품 : 두 개 이상의 상품이 참조 관계를 갖는 경우 하위(자식) 상품의 정보는 상위(부모) 상품에 존재하는 레코드이어야 함</td></tr>
</table>

품질지수 계산	○ 분자기준(A) : 상위(부모) 누락 상품 건수 ○ 분모기준(B) : 하위(자식) 상품 건수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
오류 예시	○ [지역별 카드거래현황 상품]의 "지역코드"는 [카드거래 지역 마스터 상품]의 "지역코드"에 존재하지 않음

② 필요성

데이터상품 파일은 상품 간 융복합을 위해 참조키를 정의하고 있으며 마스터파일과 참조파일이 서로 참조관계를 형성하고 있다. 참조무결성은 파일 간 연결 시 레코드 참조 누락을 방지하기 위해 다음의 항목을 진단해야 한다.

- 논리적으로 관계가 있는 파일들의 데이터는 참조 무결성을 준수하는가 (같은 형식과 범위이어야 함)
- 한 번도 참조되지 않는 값이 존재한다면 불필요한 데이터일 수 있으므로 그 사유를 알 수 있는가

③ 평가 방법

- 파일 간 관계는 관계키를 통해 연계되므로 반드시 파일 간 관계가 있는 경우 연계키를 조사해야 한다.
- 예시) “화학물질 기본정보(환경빅데이터플랫폼) 상품 - 화학물질등록번호가 관계키로 설정 / 화학물질 인체유해성 정보(환경빅데이터플랫폼) 상품 ”화학물질등록번호“를 관계키로 참조 관계가 성립되어야 함

나) 항목 형식

① 개요

[표 2-22] 항목형식 일관성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 동일한 데이터 항목이 동일한 데이터 형식을 갖는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 상품 : 여러 데이터상품의 동일데이터 관리 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 상품 : 두 개 이상의 상품에 동일한 항목이 존재할 경우 해당 항목의 값의 형식은 동일해야 함
(예시) A상품의 우편번호 는 숫자 5자리 형식 / B상품의 우편번호는 숫자3 - 숫자2자리 형식)</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 형식 불일치 건수
○ 분모기준(B) : 형식 불일치 발생 대상 상품의 전체 건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [지역별 카드거래현황 상품]의 "지역우편번호"와 [지역/업종별 카드거래현황 상품]의 "지역우편번호"의 형식이 서로 상이함</td></tr>
</table>

② 필요성

데이터상품을 활용하여 서비스 제공 시 동일한 의미의 항목은 동일한 포맷으로 일관되게 제공해야 한다. 동일한 의미의 항목이 서로 다른 파일, API에서 서로 다른 형식으로 제공될 경우 포맷 변환 시 오류가 발생할 수 있다.

③ 평가 방법

- 동일한 의미의 데이터를 담고 있는 데이터 항목을 조사
- 평가 대상은 분류, 코드값등이 대상이며 항목의 데이터 값이 동일한지 점검 실시

다) 항목 값

① 개요

[표 2-23] 항목 값 일관성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 동일한 데이터 항목이 동일한 데이터 값을 갖는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 상품 : 여러 상품의 동일 의미데이터 관리 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 상품 : 두개 이상의 상품에 동일한 항목이 존재할 경우 해당 항목의 값은 동일한 값의 목록으로 표현되어야 함
(예시) A상품의 성별구분은 남,여 / B상품의 성별구분은 남자,여자 / C상품의 성별구분은 F,M</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 값 불일치 건수
○ 분모기준(B) : 값 불일치 발생 대상 상품의 전체 건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [지역별 카드거래현황 상품]의 "고객구분" 은 "개인/법인"이고 [지역/업종별 카드거래현황 상품]의 "고객구분" 은 "개인고객/법인고객"의 값으로 값이 서로 상이함</td></tr>
</table>

② 필요성

데이터 유효성 측면에서 항목 점검 시 단일 항목 하나의 유효값 준수여부에 집중해서 점검하게 된다. 예를들어 성별구분의 경우 A상품의 성별항목은 ”남,여“ 라고 기재하고 유효값이 ”남,여“라고 되어 있을 경우 A상품의 성별 항목은 문제가 없어보인다. 그러나 B상품의 성별항목은 ”남자,여자“ 라고 기재하고 유효값이 ”남자,여자“라고 되어 있을 경우 단순항목의 유효값의 의미는 문제가 없지만 값을 활용하는 입장에서는 동일한 ”성별구분”이지만 상품마다 다른값으로 관리되는 것은 다양한 문제를 발생 시킬 수 있다.

데이터상품을 활용하여 서비스 제공 시 동일한 의미의 항목은 동일한 값으로 일관되게 제공해야 한다. 동일한 의미의 항목이 서로 다른 파일, API에서 서로 다른 값으로 제공될 경우 서로 다른 의미로 사용될 수 있다.

③ 평가 방법

- 동일한 의미의 데이터를 담고 있는 데이터 항목을 조사
- 평가 대상은 분류, 코드값등이 대상이며 항목의 데이터 값이 동일한지 점검 실시

라) 항목 관계

① 개요

[표 2-24] 항목 관계 일관성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 하나의 항목값에 따라 다른 항목값이 정해지는 경우 서로 일관되게 관리되는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 항목 : 항목간에 연관관계를 갖고 있는 항목 (예시 : 결혼여부와 결혼기념일 / 고객구분코드와 주민법인번호 등)</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 항목 : 한 개의 상품 내 항목간에 연관관계를 갖고 있는 경우 연관관계는 일관되게 관리되어 함
예시 : 고객구분코드가 "개인" 일 경우 주민법인번호 항목에는 "주민등록번호"가 입력되어야 하며, "법인" 일 경우 "법인등록번호"가 입력되어야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 연관관계 불일치 건수
○ 분모기준(B) : 전체 데이터 건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [지역별 카드거래현황 상품]의 "고객구분" 은 "개인"인데 "주민법인번호" 항목에 "법인번호"가 입력된 데이터 값 존재</td></tr>
</table>

② 필요성

대부분의 데이터품질 진단이 항목 개별 값의 유효성 검증에 중심이 맞춰져 있다. 그러나 데이터는 항상 항목 간 연관 관계를 갖고 하나의 데이터값에 따라 다른 데이터값이 영향을 받는 경우가 대부분이다. 따라서 2개 이상의 항목간에 값이 일관되게 유지되는지 지속적을 관리해야 한다.

③ 평가 방법

- 고객구분이 개인/사업자에 따라 고객번호는 주민등록번호/사업자등록번호가 입력되어야 함
- 고객구분코드, 고객번호의 일관성을 SQL을 통해 상호 비교

(5) 유일성 (세부지표 : 항목, 레코드)

데이터상품에서 유일해야 하는 항목, 레코드는 동일 데이터상품에서 유일하게 제공되어야 한다.

가) 항목 유일성

① 개요

[표 2-25] 항목 유일성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터상품의 항목이 유일해야 하나 중복이 있는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 단독 유일 : 특정 항목에 중복이 없어야 하는 항목
○ 조건 유일 : 특정 항목의 조건에 따라 해당 항목에 중복이 없어야 하는 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 단독 유일 : 상품 내 특정 단일,복합 항목에 중복이 존재할 수 없음
○ 조건 유일 : 특정 항목의 조건에 따라 상품 내 특정 단일,복합 항목에 중복이 존재할 수 없음</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 중복발생 데이터 건수
○ 분모기준(B) : 전체 데이터 건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류 예시</td><td colspan="7">○ [권역 마스터 상품]의 "권역코드"에 중복 데이터가 존재
○ [지역/권역별 카드거래현황 상품]의 "기준월" 항목이 "8월"일 경우 "권역" 항목이 유일해야 하나 중복 데이터가 존재</td></tr>
</table>

② 필요성

데이터상품은 파일 특성상 DBMS 관리데이터의 Primary Key나 Unique Index처럼 중복을 방지하는 Unique Constraints가 존재하지 않는다. 이에 따라 유일하게 관리되어야 하는 데이터의 중복이 다수 발생할 수 있으며 이러한 중복은 데이터 집계 가공에 영향을 미치므로 반드시 점검을 통해 유일하게 관리해야 한다.

③ 평가 방법

- 유일하게 관리되어야 하는 항목에 대해 GROUP BY COUNT가 1보다 큰 데이터를 조사

나) 레코드 유일성

① 개요

[표 2-26] 레코드 유일성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터상품의 레코드가 유일해야 하나 식별자 항목을 기준으로 중복이 있는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 데이터상품 레코드</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 식별자 : 레코드의 식별자(유일키)를 기준으로 레코드 중복이 없어야 함</td></tr>
<tr><td>품질지수
계산</td><td colspan="7">○ 분자기준(A) : 중복발생 레코드 건수
○ 분모기준(B) : 전체 레코드 건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
<tr><td>오류
예시</td><td colspan="7">○ [권역/월/업종별 카드거래현황 상품]의 "권역,월,업종" 식별자에 중복데이터가 존재함</td></tr>
</table>

② 필요성

데이터상품은 파일 특성상 DBMS 관리데이터의 Primary Key나 Unique Index처럼 중복을 방지하는 Unique Constraints가 존재하지 않는다. 이에 따라 유일하게 관리되어야 하는 데이터의 중복이 다수 발생할 수 있으며 이러한 중복은 데이터 집계 가공에 영향을 미치므로 반드시 점검을 통해 유일하게 관리해야 한다.

③ 평가 방법

- 유일하게 관리되어야 하는 레코드에 대해 UNIQUE COUNT가 1보다 큰 레코드를 조사

(6) 적시성 (세부지표 : 데이터제공, 응답)

구매자가 원하는 시점에 데이터상품의 제공 주기에 따른 가장 최신 데이터를 제공해야 한다.

가) 데이터 제공 적시성

① 개요

[표 2-27] 데이터 제공 적시성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터상품이 갱신주기에 따라 정상적으로 갱신되었는지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 파일 : 파일의 갱신주기가 있는 경우 UPDATE 일자 또는 체크섬 값
○ 레코드/항목 : 파일 항목에 변경일시, 생성일시와 같은 생애주기 관련 항목</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형</td><td>파일
API</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>-</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 파일 : 상품명세서에 기재된 갱신주기에 따라 파일은 갱신되어야 함
○ 레코드/항목 : 상품명세서에 기재된 갱신주기에 따라 상품 내용은 갱신되어야 함
(파일 항목에 변경일시, 생성일시와 같은 생애주기 관련 항목)</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : -
○ 분모기준(B) : -
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
○ 상품명세서의 갱신주기가 없는 경우 해당 상품은 제외</td></tr>
</table>

오류 예시	○ [지역별 카드거래현황 상품]의 갱신주기는 1개월이나 파일의 UPDATE 일자가 3개월 이전으로 되어 있음 ○ [지역별 카드거래현황 상품]의 갱신주기는 1개월이고 파일의 UPDATE 일자가 현재월로 되어 있으나 파일의 체크섬은 이전파일과 동일함 ○ [지역별 카드거래현황 상품]의 갱신주기는 1개월이나 상품의 "변경일시", "등록일시" 가 모두 3개월전 일시만 존재함

② 필요성

데이터상품은 상품명세서의 갱신주기에 맞는 최신의 데이터를 제공해야 하며 구매자의 데이터 처리 요청은 처리기한 내에 반영되어야 한다. 데이터 제공이 적시에 이루어지지 않을 경우 과거 데이터가 서비스될 수 있으며 이는 잘못된 의사결정으로 연결될 수 있다.

③ 평가 방법

- 상품명세서의 갱신주기를 기준으로 현재 등록된 상품 파일의 추가/변경 일시를 비교

나) 응답 적시성

① 개요

[표 2-28] 응답 적시성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 파일/API/서비스 호출 시 데이터 송수신의 응답이 활용 가능한 수준인지 평가함</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 파일/API/서비스 : 호출 시 회신</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>정형
비정형</td><td>파일
API
서비스</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>-</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ API : API의 평균응답속도는 데이터 300건을 기준으로 3초 이내 응답해야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 응답속도
○ 분모기준(B) : 3 Sec
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
</table>

	○ 응답속도가 60초 초과 - 0점 / 3 Sec 이내 100점 ○ 다운로드 응답속도는 다운로드 호출 시작 시점을 기준으로 함
오류 예시	○ [미세먼지 수치 API 상품] 호출 시 응답까지 20초의 시간이 소요됨.

② 필요성

데이터상품을 제공하는 데이터거래소는 안정적인 데이터 제공을 위하여 데이터상품의 다운로드, API 호출 응답속도를 사용자가 만족할만한 수준으로 제공해야 한다.

③ 평가 방법

- API 데이터상품의 경우 API 호출 후 최초 데이터의 Response Time을 측정
- 파일 데이터상품의 경우 상품 다운로드 클릭 후 다운로드가 시작되기까지의 시간을 측정

(7) 활용성 (세부지표 : 개인정보 익명성)

구매자가 데이터상품을 활용하는데 있어 법적인 문제가 발생하지 않아야 하며 만족하는 수준의 충분한 정보를 제공해야 한다.

가) 개인정보 익명성

① 개요

[표 2-29] 개인정보 익명성 정의

항 목	내 용						
지표 정의	○ 개인정보가 비식별화 정책에 따라 비식별화 되어 익명성이 보장되는지 평가함						
평가 대상	○ 개인정보 보호법에 지정된 항목을 기준으로 개인정보 비식별 조치 가이드라인에 따른 항목과 기준						
평가 요소	**상품형태**	**제공방식**	**데이터역할**	**평가방식**	**생애주기**	**도메인그룹**	**품질지수**
	정형 비정형	파일 API	제공자	정량 (시스템)	생성	-	포함

평가 기준	○ 개인정보 기본 항목 : 개인정보 항목은 비식별화 정책에 따라 비식별화 되어 있어야 함 (전화번호, 메일주소, 자택주소 등) ○ 기타 항목 : 개인정보가 포함될 수 없음 (항목에 개인정보가 포함되어 있는지 검사)
품질지수 계산	○ 분자기준(A) : 비식별 데이터 건수 ○ 분모기준(B) : 전체 데이터 건수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
오류 예시	○ [카드거래내역]의 " 이메일주소" 항목에 평문의 이메일주소가 존재함 ○ [카드거래내역]의 " 사용내역" 항목에 평문의 모바일번호가 존재함

② 필요성

유통하고자 하는 데이터에 개인정보가 포함되어 있을 경우 특정 개인이 식별되지 않도록 「개인정보 비식별 조치 가이드라인」을 준수한 비식별 조치가 필요하다.
이에 따라 개인을 식별할 수 있는 정보가 있는 경우, 이의 일부 또는 전부를 삭제하거나 일부를 개인 식별이 불가한 정보로 대체하여 다른 정보와 결합하여도 개인을 특정할 수 없도록 하는 조치해야 하며 이를 주기적으로 점검해야 한다.

③ 평가 방법

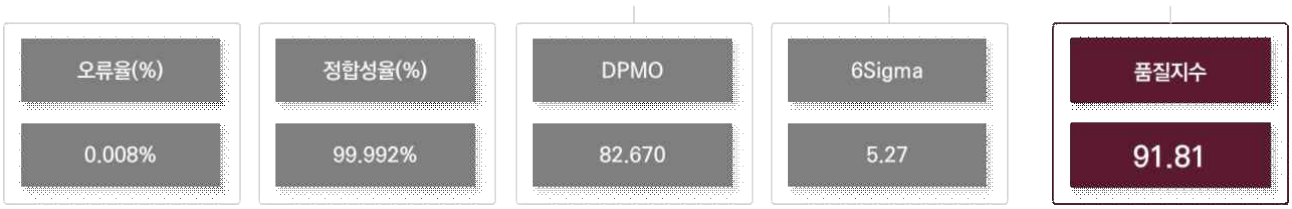
- 개인정보 항목에 해당하는 데이터가 평문 형태로 제공되고 있는지 점검

4.6. 품질평가 결과

각각의 속성별로 품질평가를 수행한 결과는 가치평가모델에 제공되어 가치를 평가하는데 활용된다. 따라서 데이터품질 평가결과는 데이터상품의 오류율(%), 정합성율(%), DPMO, 시그마, 품질지수의 형태로 다음과 같이 산출된다.
각각의 항목은 1개 이상의 품질평가지표에 대응되어 평가되며 해당 결과는 오류건수와 전체건수의 형태로 수집된다.
품질평가 결과는 항목 및 지표별로 각각 계산되며 최종 결과는 오류건수의 전체합계와 전체건수의 전체합계를 기준으로 산출한다.

구분	완전성		유효성			정확성			일관성				유일성		활용성	전체
	항목 완전성	레코드 완전성	범위 유효성	형식 유효성	목록 유효성	의미 정확성	계산/집계 정확성	선후관계 정확성	참조 무결성	항목 형식 일관성	항목 값 일관성	항목 관계 일관성	항목 유일성	레코드 유일성	개인정보 익명성	
거래일자	0		0	0										0		0
품목	0				0											0
품목종류	0				0							0				0
품목중량	0		0							3						3
중량단위	0				1,328					0						1,328
거래가격	0		0													0
상품등급	0				0					4						4
유통구분	0				0					4						4
도시	0				0											0
진단항목수	9	0	3	1	6	0	0	0	0	4	0	1	0	1	0	25
오류건수	0	0	0	0	1,328	0	0	0	0	11	0	0	0	0	0	1,339
전체건수	5,828,805	0	1,942,935	647,645	3,885,870	0	0	0	0	2,590,580	0	647,645	0	647,645	0	16,191,125
오류율(%)	0.000%		0.000%	0.000%	0.034%					0.000%		0.000%		0.000%		0.008%
정합성율(%)	100.000%		100.000%	100.000%	99.966%					100.000%		100.000%		100.000%		99.992%

[그림 2-13] 지표-항목별 데이터 품질평가 결과



[그림 2-14] 최종 데이터상품 품질지수 산정결과

PART 05

Korea Data Agency

비정형데이터

5.1. 개요

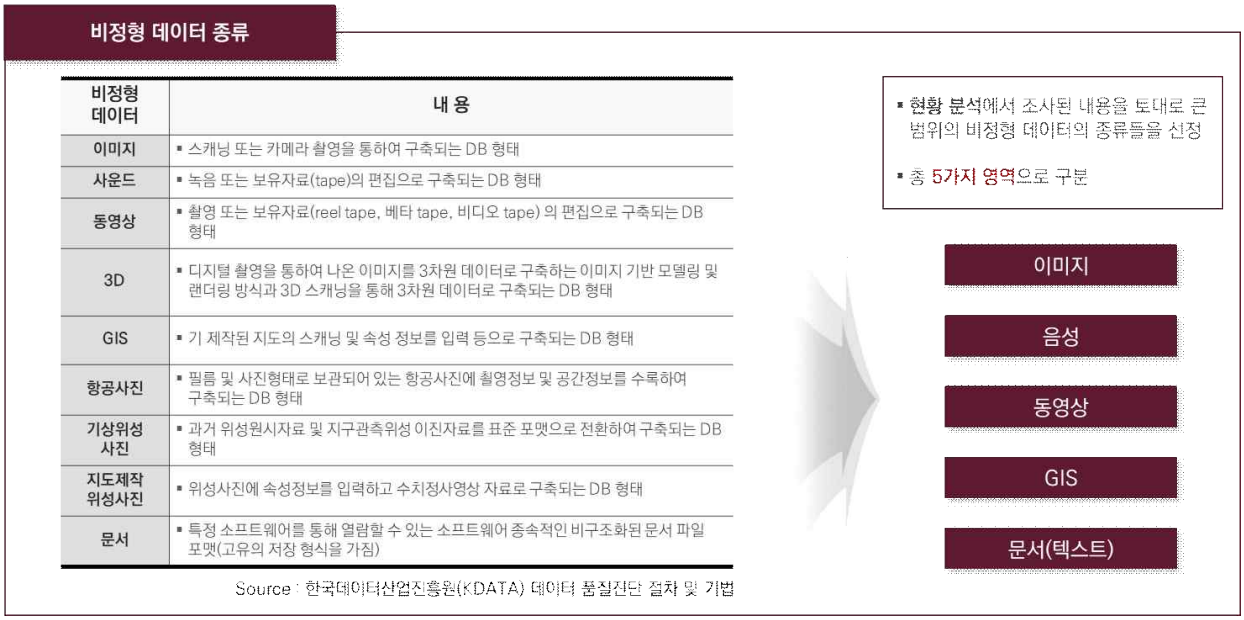
지금까지 연구된 데이터의 품질들은 공급자 입장에서 반드시 지켜야 할 내용들에 대해 기술되었다. 그렇다 보니 지켜야 할 품질의 대상을 정의하고 진단의 기준들을 정의하여 공급자가 해당 지표들에 대한 높은 품질을 유지하는 것을 유도하였다. 그러나 거래소에서 바라보는 품질의 관점은 공급자가 반드시 지켜야 할 품질이 아니라 구매자에게 데이터의 정보를 제공하는 것을 목적으로 한다. 어떤 지표에 대한 점수가 낮더라도 구매자 입장에서 해당 지표가 구매를 결정하는데 중요하지 않을 수 있으며 어떤 지표는 100% 유지 되어야 구매 결정할 수 있는 근거일 수 있을 것이다. 따라서 이번 거래소 활성화를 위한 품질 관리 가이드에서는 거래소가 평가할 수 있는 품질평가지표들에 대하여 객관적 기준 및 방법을 정의하여 품질평가에 대한 투명성을 확보하고자 한다.

정형 데이터 중심의 데이터 품질 연구는 데이터 분석보다는 단순한 비교 측정을 통한 방법들이 주를 이루었다. 그러나 데이터의 양이 방대해지고 그 종류도 다양해지면서 데이터 분석에 대한 필요성이 대두되었고, 100%의 품질 정확도가 아닌 충분(Good Enough)한 품질 수준을 요구하게 되었다. 이러한 현장의 요구 및 시대의 변화에 따라 이번 거래소 활성화를 위한 품질평가지표들에도 분석적인 요소들 나아가 인공지능(딥러닝[1]) 기술들에 대한 활용까지 포함하였다.

1) 딥러닝(Deep Learning) : 기계학습의 한 분야로써 인간의 신경망을 그대로 재현하여 구성한 고급 분석 기법

5.2. 정의

비정형 데이터란 정형 데이터 이외의 데이터를 일컬으며 비즈니스 상에서 생산되는 데이터 중 대다수를 차지하고 있을 만큼 중요한 영역으로 자리 잡고 있다. 정형화되어 있지 않은 데이터라 함은 데이터 내부적 구조는 존재하나 사전에 정의되는 데이터 모델[2]이 존재하지는 않음을 의미 한다. 비정형 데이터의 종류는 크게 이미지, 음성, 동영상, GIS[3], 문서(텍스트) 5가지로 구분한다.



[그림 2-15] 비정형 데이터 종류

각각의 비정형 데이터 종류들이 가지고 있는 특성과 분석 방법은 너무나도 다르고 그것들을 실제 구현해서 검증하는 일은 많은 시간과 노력이 동반된다. 또한 각 타입의 분석 수준도 각기 다르기 때문에 모든 비정형을 다루는 것은 불가능하다. 그래서 거래소 활성화를 위한 데이터 품질관리에서는 현재 가장 연구가 활발하게 진행되고 있으며, 기술적으로 높은 수준에 도달하고 있는 이미지를 중심으로 지표들을 정의하고 그 방법들을 제공한다.

2) 데이터 모델 : 데이터에 대한 정의 및 데이터들간의 관계와 흐름에 관한 추상화된 모형
3) GIS : Geographic Information System의 약자로 지리적 공간데이터를 표현하기 위한 시스템 또는 데이터

5.3. 대상

비정형 데이터를 바라보는 관점은 크게 두 가지 영역으로 구분된다. 첫 번째로 비정형 데이터 그 자체에 대한 단일 품질이다. 이미지 데이터의 선명도, 영상 데이터의 음성 동기화, 음성 데이터의 노이즈 소리 및 문서의 가독성 등이 단일 품질 영역에 대한 품질이다. 그러나 단일 품질에 대한 점수를 산출하는 과정은 쉽지 않다. 단순히 점수를 산출하는 것은 가능할지라도 그에 대한 품질의 높낮이를 판단하는 것은 쉽지 않은 일이다.
두 번째로 다수의 비정형의 데이터가 모여서 하나의 데이터 세트가 구성되고 이를 관리하는 관점에서의 품질이다. 특히 거래소에 등록되는 데이터들은 개별 구성보다는 다수의 데이터가 하나의 세트로 구성되는 일이 많기 때문에 해당 가이드는 두 번째의 관점으로 품질에 대한 측정 및 평가 방법을 정리하였다. 거래소 관점을 벗어나더라도 기업들은 내부의 비정형 데이터 관리를 위해서는 반드시 이러한 관리적 관점에서 바라보는 품질의 지표들이 필요하다.
데이터 세트의 관리 관점에서 보면 비정형 데이터의 평가 대상은 크게 3가지로 구분한다. 관리 메타 데이터, 객체 메타 데이터, 객체(실) 데이터로 구분하며 각각의 정의는 아래 [표2-30]과 같다.

[표 2-30] 품질 진단 관리 대상

구분	정의	예시
관리 메타 데이터	하나의 데이터 셋으로 통합되어 관리가 될 때 효율적인 활용을 위해 각 객체들에 대한 정보를 하나로 수집한 데이터	객체 메타 데이터, 데이터 개수, 데이터 주제 영역, 데이터 세트 경로
객체 메타 데이터	활용 및 관리를 위해 생성되는 객체들의 정보	파일명, 확장자, 경로, 크기
객체 데이터	순수한 객체 자체의 데이터 영역	선명도, 밝기, 노이즈, 주파수, 흔들림, GIS 라인 오버랩, 의미 단어 비중

5.4. 품질평가지표 구성

앞서 정의된 7가지의 품질평가지표 대분류를 기준으로 비정형 데이터 관련 13가지의 세부 분류 지표를 아래의 [표2-31]과 같이 정의한다. 세부 품질평가지표들은 *"NIA 공공데이터 품질 관리 매뉴얼, LX 국가공간정보 표준화 연구의 품질 지표, KDB 데이터 품질 진단 및 기법, 미래부 빅데이터 품질 진단, NIA 빅데이터 플랫폼 센터의 데이터 품질관리 가이드"*를 참고로 하여 개발되었으며, 거래소 활성화를 위한 데이터 품질관리에서 새롭게 개발한 지표들도 포함되었다.

[표 2-31] 품질 진단 세부 지표

대 분류	세부 분류	정의
완전성	메타 완전성	○ 메타데이터의 유무 및 필수 항목 내 Null 값 진단
유효성	응답 유효성	○ 데이터 파일 자체의 오류 여부를 평가함
	기능성	○ 객체 데이터가 가지고 있는 자체적인 품질에 대하여 평가함
정확성	메타 정확성	○ 관리 메타 데이터에 기재된 값과 실제 객체 메타 데이터의 값이 정확한지 평가함
	의미 정확성	○ 객체 데이터의 내용의 의미가 정확하지 않은지 평가함
	주제 정확성	○ 데이터의 주제와 실제 객체 데이터의 내용이 일치 하는지 평가함
적시성	최신성	○ 데이터가 만들어진 시점과 현 시점을 기준으로 최신의 자료인지를 평가함
일관성	메타 일관성	○ 객체 메타 데이터들에 대하여 일관되게 관리 하고 있는 가를 평가함
	메타 유사성	○ 객체 메타 데이터들에 대하여 균질한 규격을 가지고 있는지 평가함
유일성	객체 유일성	○ 객체 데이터가 다른 파일명을 가지고 중복되어 있는지 평가함
활용성	친밀성	○ 활용의 접근이 용이한 보편적인 확장자[4]를 갖는 파일인가를 평가함
	효율성	○ 상품으로서의 데이터를 활용하는 것에 있어 특성을 평가함
	개인정보 익명성	○ 개인정보가 비식별화 정책에 따라 비식별화[5] 되어 익명성이 보장되는지 평가함

4) 확장자 : 컴퓨터의 파일의 이름에서 해당 파일의 종류와 역할을 표시하기 위해 사용되는 부분
5) 비식별화 : 개인의 정체성이 들어나지 않도록 데이터를 처리하는 과정

5.5. 품질평가지표 상세

(1) 완전성 (세부지표 : 메타 완전성)

가) 메타 완전성

① 개요

[표 2-32] 메타 완전성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 상품 등록 시 등록되어야 할 메타데이터의 유무 및 필수 항목 내 Null 값 진단</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 상품명세서
○ 필수 항목의 완전성</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>비정형</td><td>파일</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 상품명세서: 비정형 데이터 세트를 관리하기 위한 정의서 파일의 유무
○ 필수 항목의 레코드 완전성 : 비정형 상품명세서의 필수 항목들의 내용에 대한 NULL 값 존재 유무</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 관리 메타 데이터의 필수 항목들의 내용이 NULL 값인 건수
○ 분모기준(B) : 관리 메타 데이터의 필수 항목들의 전체 내용 건수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수
○ 관리 메타 데이터 비존재 시 지표 점수 측정 불가 (최하점 부여)</td></tr>
</table>

② 필요성

다수의 비정형 데이터 파일을 관리하기 위해서는 반드시 관리 메타 데이터가 필요하다. 이를 통해 우리는 데이터의 기본적인 특성들을 파악할 수 있고 활용 시 필수적인 참고 데이터가 된다. 메타 완전성 지표는 관리 메타 데이터의 유무 및 필수 항목들에 대해 완전한 정보가 누락없이 입력되었는지 파악하기 위함이다.

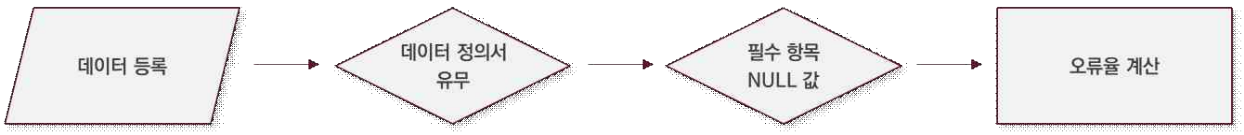
③ 평가 기준

관리 메타 데이터의 양식과 항목은 관리의 목적, 주체, 데이터의 활용에 따라 변경될 수 있다. 거래소 중심의 관리 메타 데이터의 필수 항목들은 다음과 같이 정의한다.

- 공통 - 파일명, 확장자, 파일크기
- 이미지 - 해상도
- 영상 및 음성 - 재생시간
- 문서(텍스트) - 업데이트 날짜

④ 평가 절차

평가 절차는 데이터가 등록 되면 상품명세서 유무를 확인하고, 필수 항목들에 대한 NULL 값을 측정한 뒤 오류율(%)을 산출한다.



(2) 유효성 (세부지표 : 응답 유효성, 기능성)

가) 응답 유효성

① 개요

[표 2-33] 응답 유효성 정의

항 목	내 용						
지표 정의	○ 데이터 파일 자체의 오류 여부를 평가						
평가 대상	○ 객체 데이터 파일						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정량 (시스템)	생성	-	포함

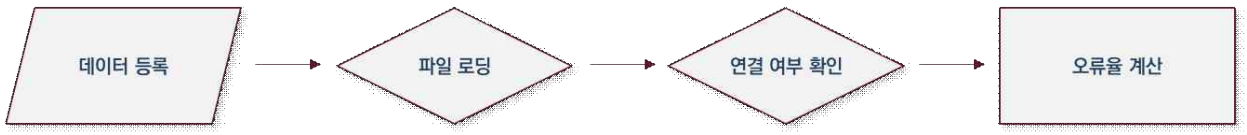
평가 기준	○ 객체 데이터 파일 : 데이터 타입과 확장자를 이용하여 데이터를 로딩 했을 시 정상적으로 파일이 동작 여부
품질지수 계산	○ 분자기준(A) : 연관된 파일 로더에서 비정상적으로 응답된 데이터 파일 개수 ○ 분모기준(B) : 실제 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수

② 필요성

비정형 데이터는 대부분 개별 파일로 저장되어 있으며, 그 내부에는 구조화된 데이터가 존재한다. 해당 파일을 읽어 들여 내부의 데이터를 파악하는 것이 품질 측정에 시작점이다. 응답 유효성은 이러한 데이터 파일 자체의 에러를 파악하기 위함이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 각 데이터 파일을 로더를 통해 로딩하게 되고 오류 여부를 판단한다. 오류 파일 개수를 확인하여 오류율(%)을 산정한다.



나) 기능성

① 개요

[표 2-34] 기능성 정의

항 목	내 용
지표 정의	○ 객체 데이터가 가지고 있는 자체적인 품질에 대하여 평가
평가 대상	○ 객체 데이터의 고유 품질 영역 1. [이미지] 선명도, 밝기 2. [영상] 선명도, 밝기 3. [음성] 주파수 4. [문서] 의미 단어의 비중, 문서 제목과의 연관성 5. [GIS] 위치 항목의 데이터 셋과 지형, 지물간의 관계

	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
평가 요소	비정형	파일	제공자	정량 (시스템)	생성	-	비 포함
평가 기준	○ 각 객체가 가진 고유 품질에 대한 측정은 가능하나 측정값에 대한 품질 평가 기준은 정의하기 어려움. 구매자의 주관적 판단을 위한 정보 제공용으로 활용 되는 지표						
품질지수 계산	○ 각각의 데이터의 고유 품질 영역을 측정한 뒤 이를 등급화하여 데이터상품의 전반적인 현황을 공유						

② 필요성

비정형 데이터는 그 종류에 따라 품질의 대상 및 측정 방법들이 다르고 측정된 값들에 품질 판단 여부도 명확하지 않다. 그러나 명확한 기준에 의해서 측정된 지수들의 경우 해당 데이터의 구매자들에게 필요한 정보를 제공할 수 있다. 기능성 지표는 이러한 정보를 제공하기 위함이다.

③ 평가 대상

평가 대상 항목에서도 언급했듯이 데이터의 종류에 따라 대상들이 상이하다. 각 평가 대상에 따라서 측정하는 방법이 다르고 적용되는 기준도 다르다. 이번 거래소 활성화를 위한 품질 관리에서는 이미지를 중심으로 평가 대상과 각 대상의 측정 방법 그리고 결과에 대한 정보화를 정의한다. 이미지의 평가 대상은 2가지 영역을 대상으로 정의한다.

- 밝기 : 이미지의 색상들의 전체적인 명암의 정도
- 선명도 : 이미지 내 픽셀과 픽셀 사이에 색상의 경계가 명확하여 피사체를 뚜렷하게 보여주는 정도

④ 평가 절차

평가 절차는 데이터가 등록되면 각 데이터 파일별로 기능성 수치들을 측정하고 정의된 범위 기준을 토대로 등급을 나눈다. 모든 데이터에 대한 측정 및 등급화를 진행하고 난 후 전체 등급 분포를 제공한다.

⑤ 평가 방법

- 밝기 : 이미지의 RGB 코드[6] 값을 추출하여 각 픽셀[7]들이 가지는 코드 값(명도) 평균을 계산한다. 그리고 255로 나눈 뒤 100을 곱하여 0~100(%)사이의 값을 갖도록 한다.
- 선명도 : 이미지를 흑백으로 변환한 뒤 라플라시안(Laplacian) 필터[8]를 적용한다. 이는 픽셀과 픽셀 간의 경계의 밝기를 더 대비적으로 표현하기 위함이다. 그리고 변환된 코드 값들의 분산을 계산한다.

⑥ 등급 기준

- 밝기 : 밝음(80~100), 보통(20~80), 어두움(0~20)
- 선명도 : 높음(1000 이상), 보통(100~1000), 낮음(0~100)
- 등급화의 목적은 구매자들이 품질을 확인하는데 쉽게 참고할 정보를 제공하고자 함이다. 등급을 나누는 범위들은 데이터의 내용이나 목적에 따라 변경된다.
- 등급을 해석하는데 주의해야 한다. 이는 단순한 수치의 높고 낮음 일뿐 품질의 척도는 아니다. 다시 말해 이미지가 어둡다 하여 품질에 문제가 있는 것은 아니다.
  이미지의 배경이 어두운 밤이라고 한다면 충분히 밝기의 수준이 낮을 수 있다.
  마찬가지로 유사한 색상의 사진들은 선명도가 낮을 수 있다.

6) RGB 코드 : 컴퓨터의 이미지 파일을 구성하는 색상 코드. 빨강(RED), 초록(GREEN), 파랑(BLUE)
7) 픽셀(Pixel) : 컴퓨터의 이미지 또는 화면을 구성하는 가장 기존 단위. 가로와 세로의 픽셀 개수를 해상도라 함.
8) 라플라시안(Laplacian) 필터 : 이미지에 라플라시안 커널을 적용하여 이미지의 윤곽선을 강조함

(3) 정확성 (세부지표 : 메타 정확성, 의미 정확성, 주제 정확성)

가) 메타 정확성

① 개요

[표 2-35] 메타 정확성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 관리 메타 데이터에 기재된 값과 실제 객체 메타 데이터의 값이 정확한지 평가</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 1. 건수 / 2. 파일명 / 3. 확장자/ 4. 파일크기 / 5. [이미지] 해상도, [영상 및 음성] 재생시간, [문서(텍스트)] 업데이트 날짜</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>비정형</td><td>파일</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 데이터건수 : 관리 메타 데이터에 입력된 건수와 실제 데이터 파일의 건수가 동일해야 함
○ 파일명 : 관리 메타 데이터에 입력된 파일명과 실제 데이터 파일명은 전부 동일해야 함
○ 확장자 : 관리 메타 데이터에 입력된 확장자와 실제 데이터 파일 확장자는 전부 동일해야 함
○ 파일크기 : 관리 메타 데이터에 입력된 파일 크기와 실제 데이터 파일 크기는 전부 동일해야 함
○ 해상도/재생시간/업데이트날짜 : 관리 메타 데이터에 입력된 각 타입별 필수 입력 메타 항목들의 정보와 실제 데이터의 정보는 전부 동일해야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A)
1. | 관리 메타 데이터에 입력 건수 - 실제 데이터 파일 전체 개수 |
2. 관리 메타 입력 값과 실제 데이터의 메타와 비 일치 개수
○ 분모기준(B) : 실제 데이터 파일 전체 개수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
</table>

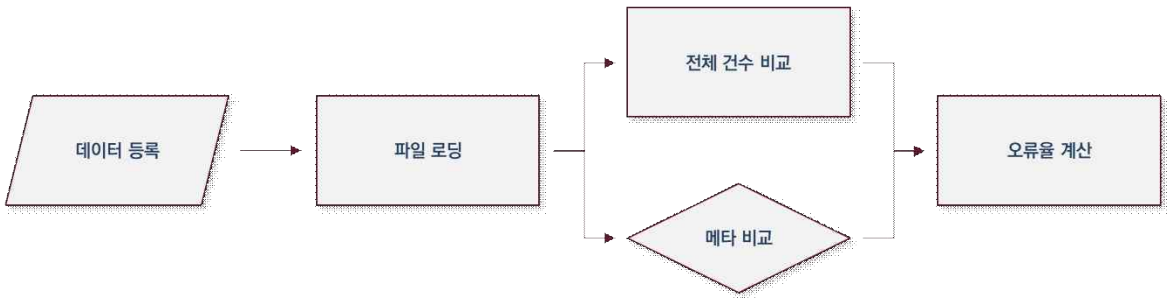
② 필요성

구매자는 데이터를 구매 후 가장 먼저 상품명세서를 통해 데이터의 내용을 파악한다. 가장 기본적인 정보부터 부가적인 내용까지 상품명세서에 적힌 내용은 해당 비정형 데이터의 정보이다. 메타 정확성은 메타 완정성에서 정의한 상품명세서 필수 입력 항목들에 대하여

실제 데이터가 가지는 값들과 동일한지 평가하기 위함이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 데이터 정의서의 등록된 건수와 실제 데이터 파일의 건수를 비교하고 동시에 정의서의 입력된 메타 데이터와 실제 데이터 파일의 메타 데이터를 비교하고 오류 여부를 파악한다.
데이터 건수의 일치율과 오류 파일 개수를 확인하여 오류율(%)을 산출한다.



나) 의미 정확성

① 개요

[표 2-36] 의미 정확성 정의

항 목	내 용						
지표 정의	○ 객체 데이터의 내용이 실제 활용가능 한 수준인가에 대한 평가						
평가 대상	○ 객체 데이터의 내용: 상품으로서의 최소한의 수준						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정량 (시스템)	생성	-	포함
평가 기준	○ 의미 파악이 불가능한 타입 별 객체 데이터의 내용 1. [이미지] 단순 바탕 또는 너무 작은 객체 2. [영상] 영상과 소리의 싱크가 맞지 않음 3. [음성] 소리가 출력 되지 않음 4. [문서] 비식별 문자로 구성 5. [GIS] 공간정보 데이터가 기준 정보와 완전 비 일치						
품질지수 계산	○ 분자기준(A) : 데이터의 의미 파악이 불가능한 데이터 파일 개수 ○ 분모기준(B) : 실제 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수						

② 필요성

파일이 오류 없이 로딩되면 그 안의 데이터는 목적에 맞게 사용하게 된다. 그러나 해당 데이터가 가지는 정보의 수준이 최소한의 기준을 넘지 못한다면 상품으로서의 품질에는 문제가 있다. 의미 정확성은 비정형 데이터가 제공하는 정보의 수준에 대하여 기준값을 넘기고 있는지를 평가하기 위함이다.

③ 평가 절차

평가 절차는 이미지 데이터를 중심으로 구성하였다. 이미지에서의 정보 주체는 바탕화면 위에서 그려지는 객체들이다. 이러한 객체가 표현되지 않거나 너무나 작게 표현되고 있다면 저품질의 데이터상품이 된다. 객체의 표현 정도를 측정하기 위해서 RGB 코드값들의 분포를 활용한다. 보통의 이미지의 RGB 코드값 분포는 넓게 퍼진 정도를 보이지만 단순 바탕이나 그 속의 작은 객체를 표현한 이미지는 좁은 RGB 코드값의 분포를 나타내고 있다. 따라서 각각의 RGB 코드들의 표준편차[9]를 활용하여 기준값(5) 이하의 값을 갖는 이미지를 오류로 파악한다. 오류 파일 개수를 확인하여 오류율(%)을 산출한다.

오류 판단 기준 표준편차 값(5)은 이미지의 특성이나 활용 목적 등에 따라 변경된다.

다) 주제 정확성

① 개요

[표 2-37] 주제 정확성 정의

항 목	내 용						
지표 정의	○ 데이터 세트의 주제 영역과 실제 객체 데이터의 내용이 일치 하는지 평가						
평가 대상	○ 객체 데이터의 컨텐츠						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정량 (시스템)	생성	-	포함

9) 표준편차 : 데이터가 가지는 퍼짐의 정도, 분산의 양의 제곱근

평가 기준	○ 객체 데이터의 컨텐츠 : 정의된 데이터의 주제 영역에 대해 객체 데이터 내에서 반드시 구현되어 있어야 함
품질지수 계산	○ 분자기준(A) : 주제에 대한 내용이 없는 데이터 파일 개수 ○ 분모기준(B) : 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수

② 필요성

개별 데이터를 하나의 데이터 세트로 구성하기 위한 기준 중에 하나로서 데이터가 표현하는 주제 영역이 있다. 그리고 주제로 선정된 내용들은 반드시 데이터 내에 표현되어야 한다. 주제 정확성은 정의된 데이터의 주제가 실제로 데이터에 포함되어 있는지를 평가하기 위함이다.

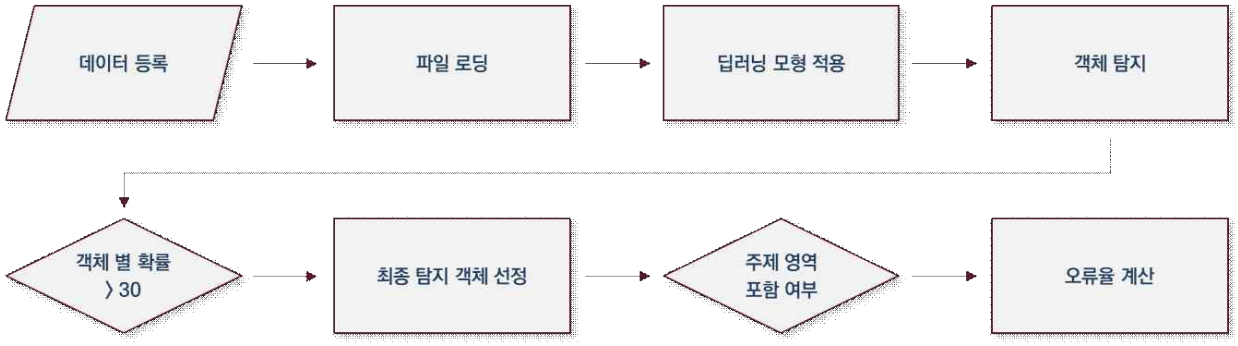
③ 인공지능 기법

이미지에서 주제의 표현은 해당 주제의 객체가 이미지에 존재하는가로 판단한다. 가장 정확한 방법은 이미지 파일을 열어 육안으로 판단하는 것이다. 하지만 이러한 방법은 엄청난 시간과 노력이 들어가는 비효율적인 방법이다. 최근 이미지 내에 객체들을 식별하고 분류하는 인공지능 기법들이 개발되어 공개되고 있다. 거래소 활성화를 위한 품질 관리에서는 이러한 인공지능 기법을 적용한 품질 측정 방법에 대해 정의하고 활용하고자 한다.

④ 평가 절차

평가 절차는 이미지 데이터를 중심으로 구성하였다. 주제 정확성을 측정하기 위하여 텐서플로우[10]에서 제공하는 객체 탐지 모형을 활용하였다. 등록된 이미지는 객체 탐지 모형을 통해 기존에 학습된 객체들을 탐지하게 된다. 탐지된 객체들은 저마다 확률값들을 함께 가지고 있는데 특정 기준 확률값(30%)을 근거로 최종 객체들을 선정한다. 선정된 객체에서 주제에 해당하는 객체의 포함 여부를 파악하여 측정한다. 이를 기반으로 오류 파일 개수를 확인하여 오류율(%)을 산출한다.

10) 텐서플로우(Tensorflow) : 구글에서 개발하고 오픈소스로 공개한 딥러닝 라이브러리



기준 확률값(30%)은 객체 탐지의 난이도, 이미지 복잡도 등에 따라 변경된다.

(4) 적시성 (세부지표 : 최신성)

가) 최신성

① 개요

[표 2-38] 최신성 정의

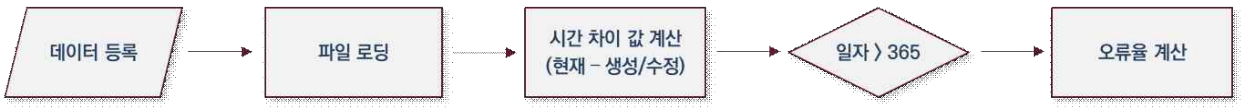
항 목	내 용						
지표 정의	○ 데이터가 만들어진 시점과 현 시점을 기준으로 최신의 자료인지를 평가						
평가 대상	○ 데이터 파일의 생성/수정 날짜						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정량 (시스템)	생성	-	포함
평가 기준	○ 데이터 파일의 생성/수정 날짜 : 데이터가 가진 정보의 내용에 따라 기준 시간 이내에 만들어진 파일이어야 함						
품질지수 계산	○ 분자기준(A) : 기준 시간 범위를 벗어난 데이터 파일 개수 ○ 분모기준(B) : 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수						

② 필요성

데이터는 현실 세계의 정보를 담고 있는 객체이다. 때로는 과거의 정보를 가지고 있는 것이 중요하고 때로는 지금의 정보를 가지고 있는 것이 중요할 수 있다. 최신성은 데이터 항목이 지금의 현실 세계 반영이라 할 때 이를 파일의 생성 및 수정 날짜를 기준으로 평가하기 위함이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 파일의 생성 또는 수정 날짜를 추출하고 현재 시점과의 시간 차이를 계산한다. 그리고 최신성 지표의 기준 날짜인 1년(365일)을 초과 여부를 통해 오류율(%)을 산출한다.



오류 판단 기준 날짜(365일)은 데이터의 특성이나 활용 목적 등에 따라 변경된다.

(5) 일관성 (세부지표 : 메타 일관성, 메타 유사성)

가) 메타 일관성

① 개요

[표 2-39] 메타 일관성 정의

항 목	내 용						
지표 정의	○ 상품에 있어서 객체 메타 데이터들에 대하여 일관되게 관리 하고 있는 가를 평가						
평가 대상	○ 파일명의 규칙성 ○ 파일 확장자의 일관성						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정량 (시스템)	생성	-	포함
평가 기준	○ 파일명의 규칙성 : 상품 내 모든 데이터 파일의 이름은 일관된 규칙성을 가져야 함 ○ 파일 확장자의 일관성 : 상품 내 모든 데이터 파일의 확장자는 일관되게 구성 되어야 함						

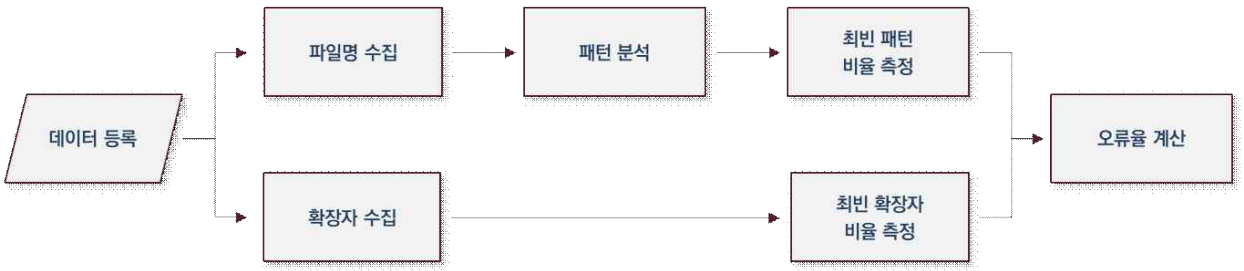
품질지수 계산	○ 분자기준(A) : 1. 파일명의 패턴 규칙을 위배한 파일 개수 2. 파일 확장자가 상이한 파일 개수 ○ 분모기준(B) : 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수

② 필요성

구매한 데이터를 활용하는데 있어서 파일들의 이름의 규칙성 및 확장자의 통일은 좀 더 쉽고 빠르게 접근을 가능하게 한다. 또한 데이터를 재가공하는 부분에 있어서도 효율적인 관리 체계를 제공한다. 메타 일관성은 데이터 파일들에 메타데이터의 통일성 여부를 평가하기 위함이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 파일들의 이름과 확장자들을 수집하고 그에 대한 통일성 여부를 측정한다. 각 패턴 중 가장 빈도가 높은 것을 기준 항목으로 정의하고 이외의 패턴에 대한 오류율(%)을 산출한다.



파일명의 규칙을 찾아내는 기준은 문자 및 숫자 조합에 대한 패턴 통일성을 확인한다.

나) 메타 유사성

① 개요

[표 2-40] 메타 유사성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 상품에 있어서 객체 메타 데이터들에 대하여 균질한 규격을 가지고 있는지 평가</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 파일크기
○ [이미지] 해상도, [영상 및 음성] 재생시간</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>비정형</td><td>파일</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 파일크기 : 상품내의 데이터 파일들의 크기는 일정 범위 내에서 균질 하게 관리 되어야 함
○ 해상도, 재생시간 : 상품내의 데이터 파일들의 해상도 및 재생시간은 일정 범위 내에서 균질 하게 관리 되어야 함
○ 일정 범위 기준 - 등록된 데이터 파일들에 대한 평가 대상 값의 평균(M)과 표준편차(S)를 이용하여 정상 범위를 구성
(M - 오류 기준 * S , M + 오류 기준 * S) / 오류 기준 : 3</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) :
1. 극단적인 파일 크기를 가지고 있는 파일 개수
2. 극단적인 이미지 해상도 또는 영상/음성의 재생시간을 가지고 있는 파일 개수
○ 분모기준(B) : 데이터 파일 전체 개수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
</table>

② 필요성

메타 일관성과 다르게 반드시 동일할 수 없지만 최소한의 유사 정도를 가져야 할 데이터 항목들이 존재한다. 하나의 데이터 항목이 다수의 것과 너무나도 큰 차이가 있다면 이는 데이터 활용에 있어서 큰 고려사항이 될 수 있다. 메타 유사성은 다수의 데이터 항목과 큰 차이를 보이는 데이터들을 찾기 위한 지표이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 파일들의 크기와 해상도를 수집하고 유사성 여부를 측정한다. 수집된 파일의 크기와 해상도의 평균과 표준편차를 계산하여 정상 범위를 정의하고 해당 범위를 벗어나는 값들에 대한 오류율(%)을 산출한다.

정상 범위를 규정하는 오류 기준 값(3)은 데이터의 특성이나 활용 목적 등에 따라 변경된다.

(6) 유일성 (세부지표 : 객체 유일성)

가) 객체 유일성

① 개요

[표 2-41] 객체 유일성 정의

항 목	내 용						
지표 정의	○ 데이터상품 내의 객체가 유일해야 하나 다른 파일명을 가지고 중복되어 있는지 평가						
평가 대상	○ 객체 데이터의 값 : 비정형 데이터를 구성하는 세부적인 데이터						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정량 (시스템)	생성	-	포함
평가 기준	○ 객체 데이터의 값 : 비정형 데이터를 구성하는 세부 데이터들 간의 중복이 없어야 함						
품질지수 계산	○ 분자기준(A) : 중복된 데이터 파일 개수 ○ 분모기준(B) : 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수						

② 필요성

파일명 중복은 대부분 시스템에서 제한하고 있는 부분이다. 따라서 파일명 중복은 거의 일어나지 않는다. 그러나 데이터 가공단계에서 발생한 에러로 인해 다른 파일명을 가지고 있지만 동일한 객체 데이터가 존재할 수 있다. 객체 유일성은 데이터를 구성하는 값을 활용하여 중복 객체를 찾기 위한 지표이다.

③ 평가 절차

평가 절차는 이미지 데이터를 중심으로 구성하였다. 이미지를 구성하는 RGB 코드를 추출한 뒤 동일 데이터 세트에 해당 코드가 정확하게 일치하는 데이터를 찾아 중복 여부를 파악한다. 이를 기반으로 중복파일 개수를 확인하여 오류율(%)을 산출한다.

(7) 활용성 (세부지표 : 친밀성, 효율성, 개인정보 익명성)

가) 친밀성

① 개요

[표 2-42] 친밀성 정의

<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 데이터를 활용함에 있어 접근이 용이하게 되어 있는가를 평가</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 데이터 파일 확장자</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>비정형</td><td>파일</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 데이터 파일 확장자 : 각 타입 별 보편적 확장자들로 구성되어 있어야 함
○ 타입 별 보편적 확장자 정의
1. [이미지] BMP, RLE, DIB ,JPG, JPEG ,PNG ,TIF, TIFF
2. [영상] WEBM, MPG, MP2, MPEG, MPE, MPV, OGG, MP4, M4P,M4V, AVI, WMV, MOV, QT, FLV, SWF
3. [음성] MP3, M4A, AAC, OGA, OGG, FLAC, PCM, WAV, AIFF</td></tr>
</table>

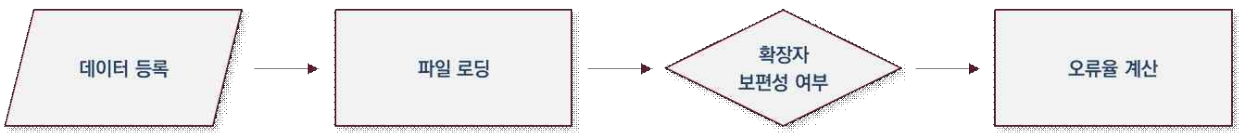
	4. [문서] DOC, DOCX, HTML, HTM, ODT, PDF, XLS, XLSX, ODS, PPT, PPTX, TXT, HWP 5. [GIS] 데이터(Shpae) 세트 구성에서 DBF, SHP, SHX, PRJ를 필수적으로 포함 (국가공간정보 표준화 문서)
품질지수 계산	○ 분자기준(A) : 비보편적 확장자를 갖는 데이터 파일 개수 ○ 분모기준(B) : 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수

② 필요성

비정형 데이터의 파일 형식은 데이터를 생산한 리소스에 따라 너무나도 다양하게 존재한다. 그 중에서도 오픈 프로그램들이 지원하면서 대중적으로 많이 사용하는 형식들이 존재한다. 데이터 구매자들은 데이터를 활용함에 있어서 보편적 파일 형식을 기대하고 추가적인 작업을 원하지 않는다. 친밀성은 이러한 추가 프로그램 설치 또는 데이터 변환 등의 활용성 측면에서 품질을 평가하기 위함이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 파일들의 확장자를 체크한다.
정의된 보편적 확장자 리스트와 비교하여 포함 여부를 확인하고 오류율(%)을 산출한다.



보편적 확장자에 대한 기준은 시간이 변함에 따라 기존의 형식들이 사라지고 새로운 것들이 일반적 형식으로 대중화될 수 있기 때문에 지속적으로 변경된다.

나) 효율성

① 개요

[표 2-43] 효율성 정의

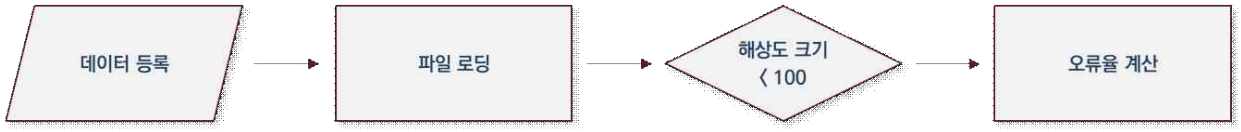
<table>
<tr><th>항 목</th><th colspan="7">내 용</th></tr>
<tr><td>지표 정의</td><td colspan="7">○ 상품으로서의 데이터를 활용하는 부분에 대한 최소한의 특성을 평가</td></tr>
<tr><td>평가 대상</td><td colspan="7">○ 타입 별 데이터 특성
1. [이미지] 해상도 크기
2. [영상 및 음성] 재생 시간
3. [문서] 가독성</td></tr>
<tr><td rowspan="2">평가 요소</td><td>상품형태</td><td>제공방식</td><td>데이터역할</td><td>평가방식</td><td>생애주기</td><td>도메인그룹</td><td>품질지수</td></tr>
<tr><td>비정형</td><td>파일</td><td>제공자</td><td>정량
(시스템)</td><td>생성</td><td>-</td><td>포함</td></tr>
<tr><td>평가 기준</td><td colspan="7">○ 타입 별 데이터 특성
1. [이미지] 해상도 크기 : 이미지들은 최소 기준 이상의 해상도를 가지고 있어야 함
2. [영상 및 음성] 재생 시간 : 영상 및 음성의 경우 최소 기준 이상의 재생 시간을 가져야 함
3. [문서] 가독성 : 문서의 경우 최소 기준 이상의 가독성을 가지고 있어야 함</td></tr>
<tr><td>품질지수 계산</td><td colspan="7">○ 분자기준(A) : 기준에 미치지 못하는 데이터 파일 개수
○ 분모기준(B) : 데이터 파일 전체 개수
○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수</td></tr>
</table>

② 필요성

데이터상품 활용 시 구매자가 기대하는 최소한의 특성 또는 성능들이 있다.
효율성은 이러한 최소 기준들에 대한 부분의 만족 여부를 평가하기 위함이다.

③ 평가 절차

평가 절차는 데이터가 등록되면 이미지의 해상도를 확인하고 기준 해상도 크기(100)와 비교한다. 기준값 이하의 이미지를 체크하여 오류율(%)을 산출한다.



기준 해상도 크기(100)는 “상품으로서 이미지 객체가 식별 가능한 크기”로 규정하였으며 데이터 특성이나 활용 목적 등에 따라 변경될 수 있다.

나) 개인정보 익명성

① 개요

[표 2-44] 개인정보 익명성 정의

항 목	내 용						
지표 정의	○ 개인정보가 비식별화 정책에 따라 비식별화 되어 익명성이 보장되는지 평가						
평가 대상	○ 개인정보의 침해를 야기 할 수 있는 항목들 1. [공통] 전화 번호, 상호명, 주소, 이메일, 주민 번호 2. [이미지, 영상] 얼굴, 자동차 번호판, 상호 간판 등						
평가 요소	상품형태	제공방식	데이터역할	평가방식	생애주기	도메인그룹	품질지수
	비정형	파일	제공자	정성	생성	-	포함
평가 기준	○ 텍스트 항목 : 객체 데이터 내에 개인 정보와 관련한 텍스트들은 비식별화 되어 있어야 함 ○ 이미지 항목 : 객체 데이터 내에 개인 정보와 관련한 이미지들은 비식별화 되어 있어야 함						
품질지수 계산	○ 분자기준(A) : 비식별 되지 않는 객체가 있는 데이터 파일 개수 ○ 분모기준(B) : 데이터 파일 전체 개수 ○ 정합성율(%) / 오류율(%) / DPMO / 시그마 / 품질지수						

② 필요성

빅 데이터 시대의 많은 비정형 데이터에서도 개인정보의 유출이 빈번하게 이루어지고 있다. 원하지 않는 나의 얼굴, 전화번호, 주소 등이 모든 데이터 타입에서 노출될 수 있다. 개인정보 익명성은 향후 문제가 될 수 있는 개인정보의 노출에 대한 평가하기 위함이다.

5.6. 품질평가 결과

정의된 평가지표의 측정을 통해 나온 결과물은 다음과 같이 제공된다.
비정형 데이터 품질평가지표는 평가결과에 대한 세부지표 점수가 의사결정에 중요한 부분이 될 수 있으므로 세부지표 점수는 다음의 결과표로 제공한다.
대분류지표 점수 또는 앞서 정의한 DPMO, 시그마, 품질지수는 필요할 경우 공식에 따라 계산될 수 있으므로 별도의 결과로 제시하지 않는다.

[표 2-45] 비정형 데이터상품 품질평가 결과표

구분	세부항목		내용				
기본 정보	데이터 세트		coco_cat				
	관리 메타 데이터		비정형정의서_cat				
	주제		고양이(cat)				
SUMMARY	파일 개수		184				
	파일 확장자		JPG		184		
	파일 평균 크기		131,956 byte				
	이미지 평균 해상도		폭	569.9	높이	471.8	
품질 점수 (정합성률)	완전성	메타 완전성	100.0				
	유효성	응답 유효성	100.0				
		기능성	밝기		선명도		
			밝음	176	상	97	
			보통	8	중	84	
			어두움	0	하	3	
	정확성	메타 정확성	개수 일치	100.0		100.0	
			값 일치	100.0			
		의미 정확성	100.0				
		주제 정확성	94.6				
	적시성	최신성	100.0				
	일관성	메타 일관성	파일명	100.0		100.0	
			확장자	100.0			
		메타 유사성	파일크기	97.8		98.9	
			해상도	100.0			
	유일성	객체 유일성	100.0				
	활용성	친밀성	100.0				
		효율성	100.0				

2020 종합안내서 데이터 거래 지원 가이드라인

발행일 2020년 12월

발행인 민 기 영

발행처 한국데이터산업진흥원

(04513) 서울시 중구 세종대로9길 42 부영빌딩 8층
TEL: 02-3708-5300 / FAX: 02-318-5040
www.kdata.or.kr / www.datastore.or.kr

본 안내서와 관련한 문의는
한국데이터산업진흥원 유통기반실 유통기획팀으로 연락해 주시기 바랍니다.

한국데이터산업진흥원 Korea Data Agency
(04513) 서울시 중구 세종대로9길 42 부영빌딩 8층
TEL: 02-3708-5300 / FAX: 02-318-5040
www.kdata.or.kr