

빅데이터 플랫폼 및 센터

2019

데이터 품질관리 가이드

제2권 데이터관리기준 | 품질진단방법

Contents

III

데이터 품질관리 기준

1장 개요	1
1. 품질관리 지표 정의	1
2. 목적 및 범위	2
2장 데이터 품질관리 지표	3
1. 지표 구성	3
2. 세부 지표	8
3. 진단 기준	10
3장 데이터 품질진단 기법	13
1. 정량적 진단 기법	13
2. 정성적 진단 기법	20
4장 성숙도 평가모델	23
1. 개요	23
2. 평가 체계	24
3. 평가기준	27
4. 평가방법	31
5. 평가항목	36

IV

데이터 품질진단 방법

1장 개요	63
1. 빅데이터 품질진단 방법 개요	63
2. 빅데이터 품질특성 및 분류체계 정의	67
3. 데이터 품질진단 방안	76
4. 품질진단 및 개선 가이드구성	88
2장 제2장 품질진단 정의(Define)	90
1. 개요	90
2. 품질관리 계획	92

3. 데이터셋(테이블) 단위 선정	93
4. 항목(컬럼) 단위 선정	99
3장 품질진단 수행(Measure)	100
1. 개요	100
2. 품질진단 수행 계획	101
3. 품질진단 규칙 정의	102
4. 품질진단 실시	142
5. 품질진단 결과	144
4장 진단결과 분석(Analyze)	145
1. 개요	145
2. 오류원인 분석	146
3. 개선기획 도출	147
5장 개선수행(Improve)	149
1. 개요	149
2. 품질개선 과제 정의	150
3. 품질개선 계획 수립	152
4. 품질개선 환경 준비	153
5. 품질개선 수행	154
6장 제6장 품질통제(Control)	156
1. 개요	156
2. 개선결과 평가	157
3. 품질목표 관리	158
4. 품질통제 실시	160



데이터 품질관리 기준

- 1장 개요
- 2장 데이터 품질관리 지표
- 3장 데이터 품질진단 기법
- 4장 성숙도 평가모델



01.

품질관리 지표 정의



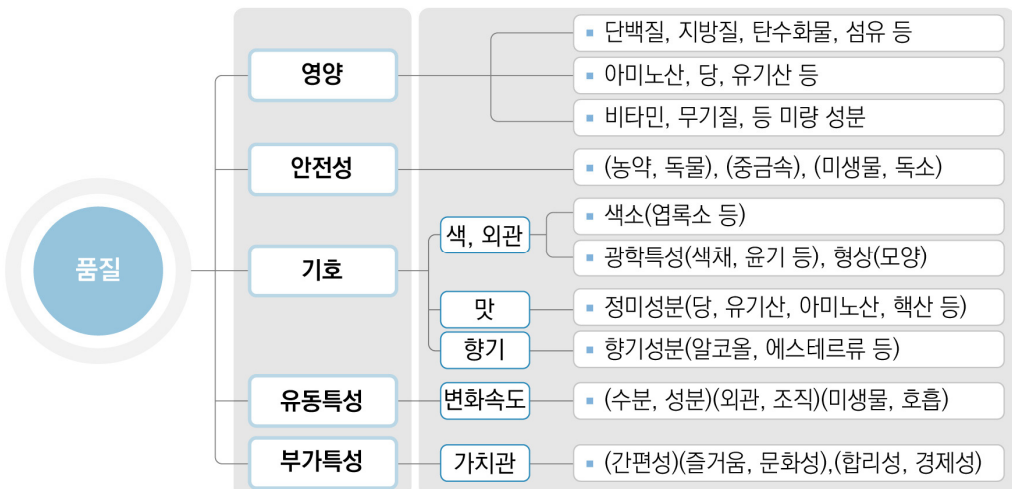
데이터의 품질을 진단하기 위해서는 사용자의 목적에 맞게 데이터가 관리되고, 사용자의 목적에 따라 어떤 것을 어떻게 진단할 것인지가 중요한 문제가 된다. 데이터는 사용자의 목적에 따른 필요 요소를 사전에 동일한 형태로 갖추고 있어야 하며, 이를 위해 데이터 품질관리 지표가 필요하다. 데이터 품질 관리 지표는 사용자의 데이터 품질 목적을 달성하는 필요 요소가 잘 갖추어졌는지 확인하기 위해 데이터 품질을 관리하는 기준이 된다.

따라서 데이터 품질관리 지표는 “데이터의 품질 수준을 측정하기 위한 관점을 정의한 것으로 무엇을 측정할 것인가에 대한 기준”이라고 정의할 수 있다.

데이터 품질관리 지표는 사용자의 목적에 맞게 데이터가 저장·관리되고 있는지 측정할 수 있는 기준이다. 지표 유형은 데이터 저장 형태와 업무적 관점의 데이터 성격, 데이터 관리 관점 등을 고려하여 정의한다. 업무규칙과 연계하여 지표별 품질 수준을 측정하고 개선 활동 수립에 활용된다.

품질지표 예시

- 식품에서 사용하는 품질지표 예시로 처음에 5가지 관점에서 품질을 측정하고 세부적으로 다양한 기준 지표, 즉 관점에서 품질을 측정하고 있음



02

목적 및 범위



기존 데이터 품질관리 기준 사례 및 빅데이터 특성 사례를 분석하여 플랫폼 및 센터 품질 관리를 강화할 수 있는 기준을 마련하고, 지속적이고 정기적인 품질 관리가 가능할 수 있도록 한다.

그리고 다양한 데이터를 수집·가공·활용하는 특성을 반영하여 범용적이고 공통적인 지표로 구성하며, 빅데이터 요소를 고려하여 정형 데이터, 비정형 데이터 관점의 기준을 포함한다.

▶ 데이터 품질관리 수립 방향성

- 빅데이터 플랫폼 및 센터 데이터 품질 강화를 위한 기준 수립
- 지속 가능한 품질관리 기준 수립
- 빅데이터 특성을 고려한 기준 반영

▶ 데이터 품질관리 기준 수립 절차

데이터 품질관리 지표는 범용적이고 공통적인 지표 구성을 위해 다양한 사례를 분석하여 기준 수립의 원칙을 정의하고, 품질관리 지표를 수립한다. 품질관리 지표별 진단 기준을 정의하고 진단 기준별 품질진단 방법을 수립한다. 이후 품질관리 지표에 기반을 둔 성숙도 평가를 위한 성숙도 모델을 수립하는 절차로 구성된다.

그림 1 데이터 품질관리 기준 수립 절차



▶ 데이터 품질관리 기준 수립 범위

데이터 품질관리 기준 범위는 빅데이터 플랫폼 및 센터에서 수집하고, 가공·분석, 활용하는 데이터를 대상으로 하며, 데이터 품질관리 지표와 세부 지표, 세부 지표별 품질 진단 기준, 지표별 품질 진단 방법, 그리고 품질수준을 측정하기 위한 데이터 품질관리 성숙도 평가 모델을 수립한다.

01

지표 구성



1.1 품질관리 지표 분석 대상

데이터 품질관리 지표는 국내·외 데이터 품질관리 및 빅데이터 품질관리 지표 등 17개 사례의 117개 지표 항목을 대상으로 지표 항목을 분석하였다.

- ▶ 국내 데이터 품질관리 지표 사례 : “공공데이터 품질관리 지표(NIA)”, “데이터 품질관리 지표(K-DATA)”, “데이터 품질관리 업무 매뉴얼(특허청)”, “데이터 품질관리 지표(국민연금)”, “데이터 품질관리 지표(건강보험공단)”, “데이터 품질 특성지표(한국정보통신기술협회)”
- ▶ 해외 데이터 품질관리 지표 사례 : “Information Quality Framework (캐나다 보건협회-CIHI)”, “Data Quality Framework (영국은행)”, “Open Data & Metadata Quality (Open Data Support)”
- ▶ 국내 빅데이터 품질관리 지표 사례 : “비정형 데이터 유형별 품질 지표 (K-DATA)”, “국가공간정보센터 품질 기준 (국토교통부)”, “소셜 빅데이터 (통계청)”, “실무자를 위한 빅데이터 업무절차 및 기술활용 (NIA)”, “비정형 데이터 품질관리 방안 연구 (NIA)”
- ▶ 해외 빅데이터 품질관리 지표 사례 : “A Suggested Framework for the Quality of Big Data (UNECE)”, “On the Meaningfulness of big data quality(Data Science and Engineering)”, “Data quality for big data (Gartner)”

I

II

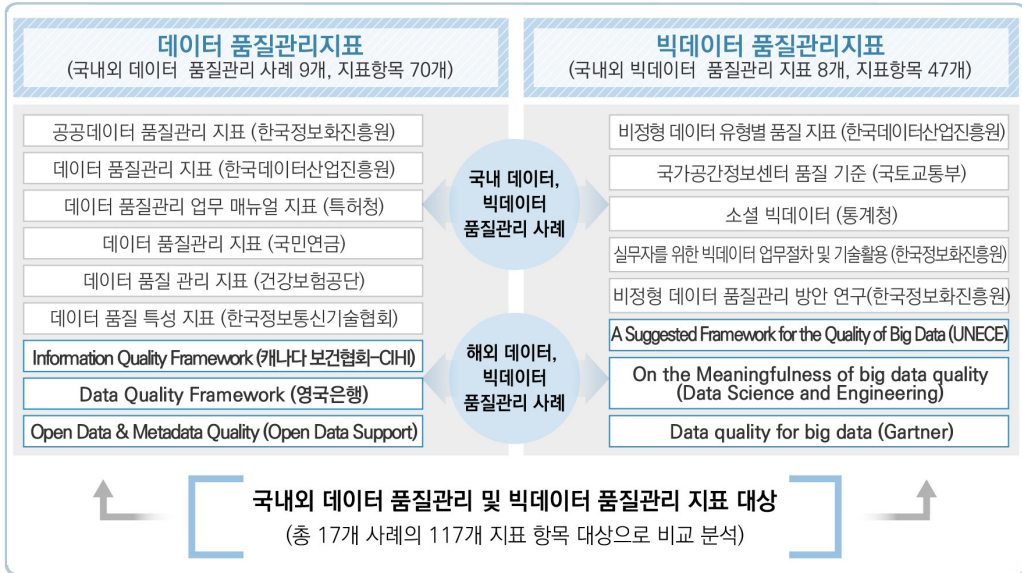
III

데이터 품질관리 기준

IV

V

그림 2 데이터 품질관리 지표 분석 대상

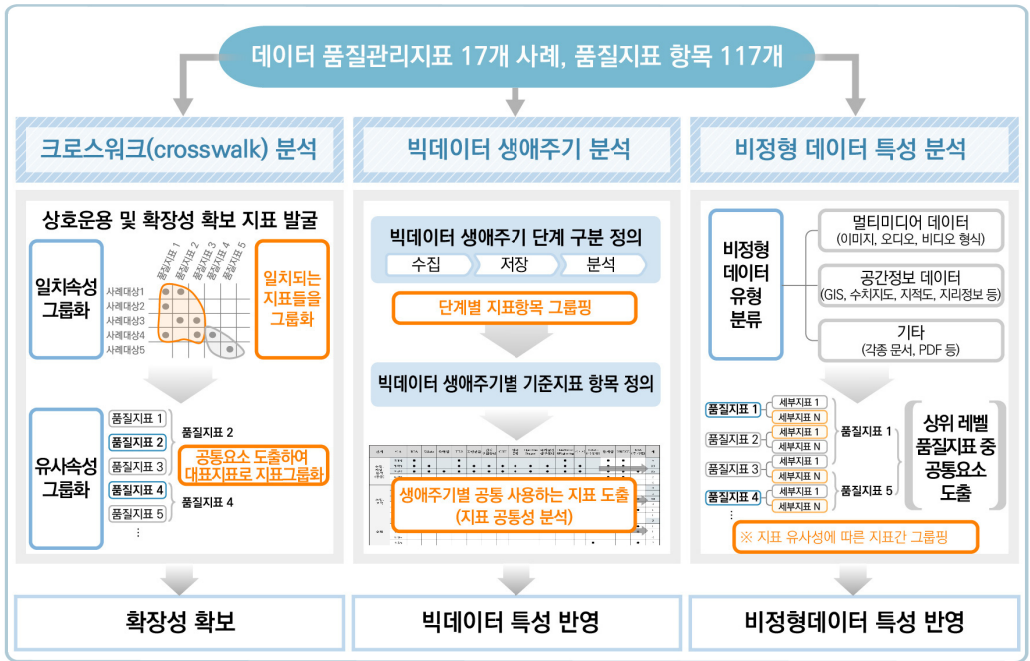


1.2 품질관리 지표 분석 방법

데이터 품질관리 지표는 보편성과 공통성을 확보하기 위해 상호운용 및 확장성 확보에 유용한 크로스 워크 방법론을 적용하여 지표를 분석하였으며, 빅데이터 특성을 반영하기 위해 빅데이터 생애주기 분석을 수행하였고, 비정형 데이터 특성을 반영하기 위해 비정형 데이터 유형을 분류하여 지표 유사성에 따른 그룹핑 방법으로 지표를 분석하였다.

- ▶ ‘크로스워크(CrossWalk) 방법’은 여러 데이터 간 데이터 요소의 의미와 구조를 매핑하는 방법으로 다양한 데이터를 대표할 수 있는 항목을 도출하여 데이터의 상호 운영, 확장성을 확보하기 위한 대표적인 방법이다.
- ▶ ‘빅데이터 생애주기 분석’에서는 생애주기 단계 및 지표 기준을 구성하고, 생애주기 단계별로 공통성 분석 및 속성 그룹화를 통해 공통 지표를 도출하여 분석한다.
- ▶ ‘비정형 데이터 분석’은 비정형데이터 유형에 따라 국내 유관기관의 비정형 데이터 품질 지표를 추출하여, 지표 간 유사성에 따라 그룹화하여 분석한다.

그림 3 데이터 품질관리 지표 분석 방법



1.3 품질관리 지표 도출

빅데이터 플랫폼 및 센터의 품질관리 지표는 크로스워크 분석, 빅데이터 생애주기 분석, 비정형 데이터 특성 분석을 통해 보편성과 공통성을 확보한 완전성, 유효성, 일관성, 정확성, 준비성, 보안성, 유용성, 접근성, 적시성 등 9가지 지표로 구성한다.

- ▶ ‘크로스워크 분석’을 통해 도출된 품질 관리 지표는 완전성, 유효성, 일관성, 정확성, 준비성, 보안성, 유용성, 접근성, 적시성 등 9개 지표로 다양한 도메인의 빅데이터 플랫폼에 적용 가능한 범용성과 확장성이 고려되었다.
- ▶ ‘빅데이터 생애주기 분석’을 통해 도출된 품질관리 지표는 완전성, 일관성, 정확성, 보안성, 유용성, 접근성, 적시성 등 7개 지표로 빅데이터 생애주기 단계에서 공통으로 사용될 수 있는 지표로 도출하였다.
- ▶ ‘비정형 데이터 특성 분석’을 통해 도출된 품질관리 지표는 완전성, 일관성, 정확성, 적시성, 신뢰성, 추적성, 타당성, 복잡성, 이식성 등 9개 지표로 비정형 데이터 유형에 따라 범용적으로 적용 가능한 지표로 도출하였다.
- ▶ 3가지 분석으로 도출된 지표에서 일치 또는 유사 항목을 그룹핑하여, 다른 지표를 포괄하여 표현할 수 있는 가장 상위 수준의 지표 항목을 데이터 품질관리 지표로 선정하였다.

그림 4 빅데이터 플랫폼 및 센터 데이터 품질관리 지표

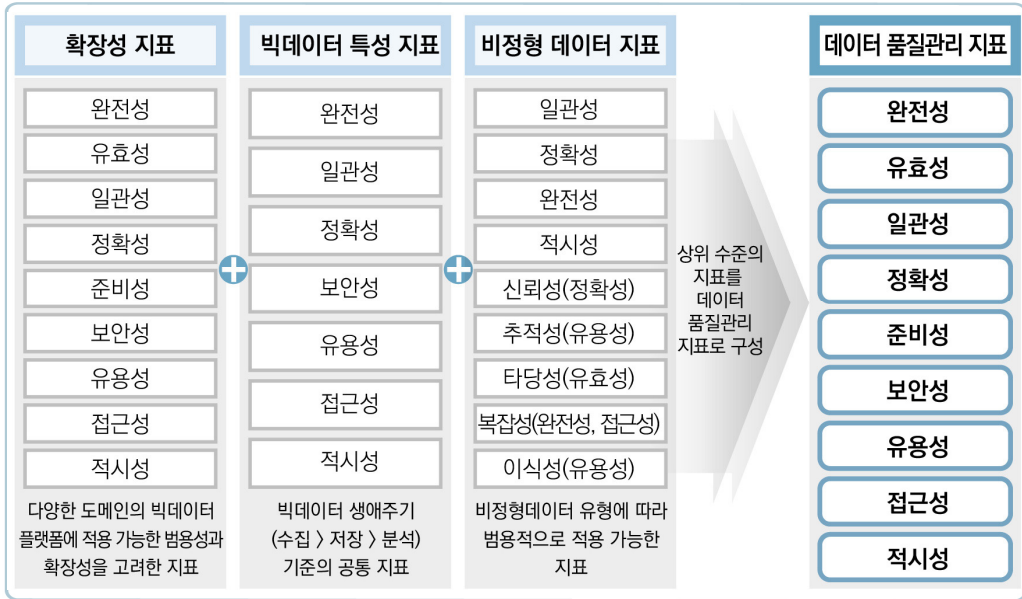


표 1 품질관리 지표 정의

지 표	정 의
완전성 (Completeness)	• 데이터의 저장소인 데이터베이스를 구축함에 있어 논리적인 설계와 물리적인 구조를 갖추고, 업무 요건에 맞게 데이터가 저장되도록 설계·구축 되었는지를 진단하는 지표
유효성 (Validity)	• 데이터가 의도한 범위와 형식으로 구성되어 해당 범위와 형식에 벗어나지 않고 본래의 효과를 제공하며 추적 가능한지 진단하는 지표
일관성 (Consistency)	• 같은 의미를 갖는 데이터는 표준을 준수하여 동일한 용어와 도메인으로 정의하고 중복, 연계 데이터의 값이 서로 일치하는지 진단하는 지표
정확성 (Accuracy)	• 데이터가 실세계 사실 및 업무규칙에 맞게 정확한 값이 저장되어 있는지를 진단하는 지표
준비성 (Preparation)	• 데이터 품질관리를 위해 기본적으로 관리해야 하는 정책, 규정, 조직, 절차 등을 마련하고, 최신의 내용으로 충실하게 관리되는지를 진단하는 지표
보안성 (Security)	• 데이터 관리 주체가 관리되고 외부 및 내부 요인으로부터 데이터를 보호하기 위해 접근이 적절히 통제되며 개인정보 등 주요 데이터에 대하여 보호 조치가 이루어져 있는지 진단하는 지표
유용성 (Usefulness)	• 사용자가 만족하는 수준의 충분한 정보가 수집되고 제공되는지, 사용자의 정보 이용에 만족도를 충족하는지, 데이터의 범위와 상세화 정도를 충족시키는지 진단하는 지표
접근성 (Accessibility)	• 사용자가 원하는 데이터를 손쉽게 사용할 수 있는지 진단하는 지표
적시성 (Timeliness)	• 사용자가 원하는 시점에 데이터의 주기에 따른 가장 최신 데이터를 유지하는지 진단하는 지표

1.4 품질관리 지표 특성

수립된 데이터 품질관리의 9개 지표는 데이터 품질관리 기반에서 활용까지 사용하여 데이터 품질 수준을 확보한다.

그림 5 품질관리 지표 간의 관계



- ▶ 데이터 품질관리 지표 준비성, 보안성은 데이터 품질관리의 기반에서 사용되는 주요 지표로 준비성과 보안성 지표의 개선은 품질관리 프로세스 개선 및 품질관리 구성 요소의 규격화를 달성한다.
- ▶ 데이터 품질관리 지표 완전성, 유효성, 일관성, 정확성은 데이터 값의 품질을 확보하는 주요 지표로 완전성, 유효성, 일관성, 정확성 지표의 개선은 데이터 값, 형식 정확성을 높이고, 데이터 연계·통합을 촉진한다.
- ▶ 데이터 품질관리 지표 유용성, 접근성, 적시성은 데이터 활용에서 사용되는 주요 지표로 유용성, 접근성, 적시성 지표의 개선은 사용자 만족도를 높이고, 데이터 서비스의 지속적인 품질을 제고한다.

I

II

III

데이터 품질관리 기준

IV

V

02

세부 지표



2.1 품질관리 세부지표 정의

빅데이터 플랫폼 및 센터의 데이터 품질관리 세부지표는 품질진단이 가능한 품질관리 대상을 값, 구조, 표준, 관리체계, 공간정보로 판단하고 품질관리 대상을 구체적으로 진단하기 위해 품질관리 지표를 세분화하여 구성한다.

표 2 품질관리 세부지표 정의

지표	세부지표	정의	대상
완전성	단독 완전성	• 데이터가 업무 요건에 맞게 항상 채워져 있는지 진단하는 지표	값
	조건 완전성	• 데이터가 다른 데이터 항목의 업무 조건에 따라 항상 채워져 있는지 진단하는 지표	값
	구조 완전성	• 데이터모델이 데이터구조 관리 지침을 준수하고 있는지 진단하는 지표 (데이터 모델과 DB와 일치 여부, 중복 테이블 존재 여부 등)	구조
	공간객체 완전성	• 공간객체가 누락되어 있거나, 초과하는 객체들이 존재하는지 진단하는 지표	공간정보
유효성	범위 유효성	• 값이 특정 리스트 내에 존재하여야 하거나, 정해진 범위 내에 존재하는지 진단하는 지표	값
	날짜 유효성	• 항목이 일자일 경우 값은 유효한 일시의 값을 갖는지 진단하는 지표	값
	형식 유효성	• 정해진 형식(Format)에 맞게 값이 존재하고 있는지 진단하는 지표	값
	코드 유효성	• 표준코드에 정의된 유효 코드 값을 갖는지 진단하는 지표	값
	여부 유효성	• 정해진 여부 기준에 정의된 유효 값을 갖는지 진단하는 지표	값
	선후관계 유효성	• 복수의 컬럼 값이 선후 관계에 있을 경우 발생 업무규칙을 지키는지 진단하는 지표	값
	그리드 유효성	• 공간 정보의 공간적 위치(격자, 좌표계) 범위가 일치하는지를 진단하는 지표	공간정보

지표	세부지표	정의	대상
일관성	참조 무결성	• FK 관계에 있는 테이블 간 참조무결성이 유지되는지 진단하는 지표	값
	중복 일관성	• 업무 목적에 따라 중복 데이터를 생성한 경우 각각의 중복 컬럼 값이 일치하는지 진단하는 지표	값
	표준 일관성	• 동일한 데이터가 표준을 준수하여 형식(도메인)과 명칭(표준용어)이 일치하는지, 엔터티 및 테이블 명명규칙을 준수하였는지 진단하는 지표	표준
	위상구조 일관성	• 공간객체가 기하학적 위상을 준수하여 생성되었는지를 진단하는 지표	공간정보
	개념적 일관성	• 공간객체간의 위상관계 규칙을 준수하고 있는지를 진단하는 지표	공간정보
	연계 일관성	• 원천 테이블과 연계 목적의 복제 테이블 간 연계 대상이 설정에 따라 같은 값으로 존재하는지 진단하는 지표	값
정확성	논리관계 정확성	• 데이터 값이 규정된 업무규칙을 준수하여 정확한 값을 유지하는지 진단하는 지표	값
	파생항목 정확성	• 파생되어 관리되는 항목에 대해 원천 항목의 값과 파생 된 항목의 값이 정확하게 관리되고 있는지 진단하는 지표	값
	메타 정확성	• 이미지, 동영상 등 비정형 데이터의 메타 속성데이터가 실제 콘텐츠의 정보와 일치하는지 진단하는 지표	값
	위치식별 정확성	• 기본공간정보를 기준으로 공간객체의 위치식별자가 일치하는지를 진단하는 지표	공간정보
준비성	체계 구비성	• 데이터 품질 관리를 위한 기본적인 요소들이 체계적으로 수립되어 관리되는지 진단하는 지표	관리체계
	체계 준수성	• 수립된 데이터 품질관리 기준 요소들이 체계적으로 수행하는지 진단하는 지표	관리체계
보안성	정보 보호성	• 비식별화 조치 등 민감정보 보호를 위한 절차와 규정을 수립하고 수행하고 있는지 진단하는 지표	관리체계
	통제성	• 외부의 침입 및 정보 훼손을 방지하는 절차와 규정을 수립하고 수행하고 있는지 진단하는 지표	관리체계
	데이터 관리 책임성	• 특정 사용자에게 데이터 오너십(접근 권한 및 책임)이 부여되고 관리되고 있는지 진단하는 지표	관리체계
유용성	충분성	• 사용자의 요구 사항을 충분히 충족시킬 수 있는지 진단하는 지표	관리체계
	유연성	• 사용자의 요구 사항을 수용할 수 있는 유연한 구조를 가지고 있는지 진단하는 지표	관리체계
	추적성	• 데이터의 메타데이터 속성으로 변경 내역이 관리되고 있는지 진단하는 지표	관리체계
	활용성	• 실제 공급되는 데이터를 사용자가 얼마나 사용하고 있는지 진단하는 지표	관리체계
접근성	검색 용이성	• 사용자가 데이터 검색할 때 보유 데이터에 대한 이해가 용이하게 제공되는지를 진단하는 지표	관리체계
	사용 편의성	• 사용자의 다양한 요구사항을 만족시킬 수 있는 서비스 및 데이터 유형이 제공되고 있는지 진단하는 지표	관리체계
적시성	데이터 최신성	• 갱신주기가 있는 배치 데이터의 수집, 갱신 주기에 따라 최신의 값을 주기적으로 갱신하고 있는지 여부를 진단하는 지표	관리체계

03

진단 기준



품질관리 진단기준은 품질관리 지표를 기반으로 하여 품질관리 진단 수행 시 진단을 수행하는 담당자가 어떻게 진단을 할 것인지에 대한 기준을 제시한 것으로 품질관리 세부지표별로 진단 기준을 구성하였다.

표 3 품질관리 진단 기준 정의

지표	세부지표	진단기준
완전성	단독 완전성	• 데이터 업무 요건에 맞게 항상 채워져 있는지 진단
	조건 완전성	• 데이터가 다른 데이터 항목의 조건에 따라 해당 조건을 충족할 때 값이 채워져 있는지 진단 (Conditional Not Null)
	구조 완전성	• 데이터 모델과 실제DB의 현행화가 잘 준수되고 있는지, 테이블 구조가 동일한 중복 테이블이 존재하는지 진단
	공간객체 완전성	• 공간객체가 누락되어 있거나, 초과하는 객체들이 존재하는지 진단
유효성	범위 유효성	• 값이 특정 리스트 내에 존재하여야 하거나, 정해진 범위 내에 존재하는지 진단
	날짜 유효성	• 값이 특정 리스트 내에 존재하여야 하거나, 정해진 유효 범위 내에 존재하는지 진단
	형식 유효성	• 데이터 값이 해당 도메인의 형식(Pattern)을 준수하는지 진단
	코드 유효성	• 표준코드에 정의된 유효 코드 값과 일치하는지 진단
	여부 유효성	• 정해진 여부 기준에 정의된 유효 값을 갖는지 진단
	선후관계 유효성	• 단일 테이블 내에서 시계열적 선후관계가 업무규칙으로 정의된 경우 발생 업무 규칙을 준수하는지 진단
일관성	그리드 유효성	• 공간 정보의 공간적 위치(격자, 좌표계) 범위가 일치하는지를 진단
	참조 무결성	• 논리적으로 관계가 있는 테이블들의 속성이 같은 형식과 범위, 값을 가지고 있는지 진단
	중복 일관성	• 동일한 데이터를 중복하여 관리하는 경우 원천데이터와 중복데이터를 구분할 수 있고, 동일한 시점에 중복데이터가 같은 값을 가지고 있는지 진단
	표준 일관성	• 동일한 데이터가 표준을 준수하여 형식(도메인)과 명칭(표준용어)이 일치하는지, 엔티티 및 테이블 명명규칙을 준수하였는지 진단

지표	세부지표	진단기준
	위상구조 일관성	• 공간객체가 기하학적 위상을 준수하여 생성되었는지를 진단
	개념적 일관성	• 공간객체간의 위상관계 규칙을 준수하고 있는지를 진단
	연계 일관성	• 원천 테이블과 연계 목적의 복제 테이블 간 연계 대상이 설정에 따라 같은 값으로 존재하는지 진단
정확성	논리관계정확성	• 데이터 값이 규정된 업무규칙을 준수하여 정확한 값을 유지하는지 진단 예) 사칙연산, 논리연산의 결과 데이터 값의 정확성 진단
	파생항목 정확성	• 파생되어 관리되는 항목에 대해 원천 항목의 값과 파생 된 항목의 값이 정확하게 관리되고 있는지 진단
	메타 정확성	• 이미지, 동영상 등 비정형 데이터의 메타 속성데이터가 실제 콘텐츠의 정보와 일치하는지 진단
	위치식별 정확성	• 기본공간정보를 기준으로 공간객체의 위치식별자가 일치하는지를 진단
준비성	체계 구비성	• 데이터 관리를 위한 정책, 규정, 지침, 절차 등이 있는지 진단
	체계 준수성	• 데이터 관리 규정이나 지침 등에 따라 데이터 품질관리 활동을 수행하고 있는지 진단
보안성	정보 보호성	• 비식별화 조치 등 민감정보 보호를 위한 절차와 규정을 수립하고 수행하고 있는지 진단
	통제성	• 외부의 침입 및 정보 훼손을 방지하는 절차와 규정을 수립하고 규정에 따라 수행하고 있는지 진단
	데이터관리 책임성	• 특정 사용자에게 데이터 오너십(접근 권한 및 책임)이 부여되고 관리되고 있는지 진단
유용성	충분성	• 제공하는 정보가 사용자의 요구사항을 충족시킬 수 있을 만큼 활용하기에 충분한 양으로 수집되는지 진단
	유연성	• 사용자의 요구 사항을 수용할 수 있는 유연한 구조를 가지고 있는지 진단
	추적성	• 데이터의 메타데이터 속성으로 데이터 변경 내역이 관리되고 데이터 변경 사항이 이력 테이블로 관리되는지 진단
	활용성	• 실제 공급되는 데이터를 사용자가 얼마나 사용하고 있는지 진단
접근성	검색 용이성	• 사용자가 원하는 데이터를 검색할 때 데이터 로딩속도, 원하는 데이터 획득, 쉬운 데이터 접근 등 보유 데이터의 접근이 용이하도록 제공되는지를 진단
	사용 편의성	• 사용자의 다양한 요구사항을 만족시킬 수 있는 서비스 및 데이터 유형이 제공되고 있는지 진단
적시성	데이터 최신성	• 갱신주기가 있는 배치 데이터의 수집, 갱신 주기에 따라 최신의 값을 주기적으로 갱신하고 있는지 여부를 진단 예) 고객마스터 테이블의 최종 고객 상태 정보는 고객이력 테이블의 마지막 데이터와 일치

I

II

III

데이터 품질관리 기준

IV

V

데이터 품질진단 방법은 세부지표를 기준으로 진단 대상별로 정량적 진단과 정성적 진단을 수행하며, 세부지표와 진단대상 값, 구조, 표준, 공간정보, 관리체계 간 매핑을 통해 해당 지표별 진단 대상만 진단하고 매핑된 진단 기법을 사용한다.

지표		진단대상					진단 기법		
		값	구조	표준	공간 정보	관리 체계	정량적 진단		정성적 진단 체크리스트
							프로파일링	업무규칙	
완전성 (Completeness)	단독 완전성	●						√	√
	조건 완전성	●						√	
	구조 완전성		●				√		
	공간객체 완전성				●		√		
유효성 (Validity)	범위 유효성	●					√		√
	날짜 유효성	●					√		
	형식 유효성	●					√		
	코드 유효성	●					√		
	선후관계 유효성	●						√	
	그리드 유효성				●		√		
일관성 (Consistency)	참조 무결성	●					√	√	√
	중복 일관성	●					√	√	
	표준 일관성			●			√		
	위상구조 일관성				●		√		
	개념적 일관성				●		√		
	연계 일관성	●					√		
정확성 (Accuracy)	논리관계 정확성	●						√	√
	파생항목 정확성	●						√	
	메타 정확성	●					√		
	위치식별 정확성				●		√		
준비성 (Preparation)	체계 구비성					●			√
	체계 준수성					●			
보안성 (Security)	정보 보호성					●			√
	통제성					●			
	데이터관리 책임성					●			
유용성 (Usefulness)	충분성					●			√
	유연성					●			
	추적성					●			
	활용성					●			
접근성 (Accessibility)	검색 용이성					●			√
	사용 편의성					●			
적시성 (Timeliness)	데이터 최신성					●			√

01

정량적 진단 기법



데이터 품질지표 정량적 진단 방법은 진단 결과가 정량적으로 도출될 수 있는 진단 대상이 값, 표준, 구조, 공간정보인 완전성, 유효성, 일관성, 정확성 지표가 대상이며 진단 내용을 기반으로 해당 지표의 진단을 수행한다.

1.1 완전성

세부지표	진단내용	분류	진단기준 및 진단방법		
단독 완전성	Null값 진단	진단상세내용	<ul style="list-style-type: none"> 데이터 업무 요건에 맞게 항상 채워져 있는지 진단 		
		진단대상	값	진단기법	업무규칙(BR)
		진단방법	<ul style="list-style-type: none"> 업무기준은 Null을 허용하지 않지만 데이터 값이 Null 인 컬럼. (업무기준은 해당 컬럼이 Not Null 정의되어야 하나 모델설계 또는 테이블 생성 시 오류로 해당 컬럼이 Null 허용 컬럼으로 정의된 경우 Null 가 발생했는지 진단. 이러한 경우 데이터 값의 개선 후 반드시 데이터 모델과 DB의 개선을 통하여 Null값이 근본적으로 발생되지 않도록 해야 함) (이러한 경우 데이터 값의 개선 후 근본적으로 해당 컬럼을 NotNull 로 변경하여 Null 값 발생을 방지) 		
		오류율 계산	$\text{오류율}(\%) = (\text{진단대상 컬럼 전체 오류 데이터 건수} / \text{진단대상 컬럼의 전체 데이터 건수}) \times 100\%$		
조건 완전성	Null값 진단	진단상세내용	<ul style="list-style-type: none"> 데이터가 다른 데이터 항목의 업무 조건에 따라 항상 채워져 있는지 진단 		
		진단대상	값	진단기법	업무규칙(BR)
		진단방법	<ul style="list-style-type: none"> 데이터 값이 특정 업무 요건에 충족하여 값이 채워져 있어야 하지만 데이터 값이 Null인 경우 		
		오류율 계산	$\text{오류율}(\%) = (\text{진단대상 컬럼 전체 오류 데이터 건수} / \text{진단대상 컬럼의 전체 데이터 건수}) \times 100\%$		
구조 완전성	테이블 중복 진단	진단상세내용	<ul style="list-style-type: none"> 전체 테이블 중에서 테이블 구조가 동일한 중복 테이블을 진단 (테이블 중복) 		
		진단대상	구조 > 테이블	진단기법	프로파일링

I

II

III

데이터 품질관리 기준

IV

V

세부지표	진단내용	분류	진단기준 및 진단방법		
		진단방법	<ul style="list-style-type: none"> 전체 테이블 중에서 테이블의 컬럼명, 데이터 타입, 데이터 길이가 100% 동일한 중복 테이블 		
		오류율 계산	<ul style="list-style-type: none"> 중복률(%) = (중복 테이블 수) / (대상DB 전체 테이블 수) x 100% 		
	테이블 현행화 진단	진단상세내용	<ul style="list-style-type: none"> 데이터 모델과 실제 DB에 구현된 테이블 불일치 정도를 진단 		
		진단대상	구조 > 테이블	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 데이터모델에 정의되지 않았으나 실제DB에만 존재하는 경우, 데이터모델에 정의되었으나 실제DB에는 존재하지 않는 경우의 불일치 테이블 		
		오류율 계산	<ul style="list-style-type: none"> 불일치율(%) = (데이터모델기준 누락 테이블 수 + 실제DB 기준 누락 테이블 수) / (데이터모델 전체 테이블 수 + 실제DB기준 전체 테이블 수) x 100% 		
	컬럼 현행화 진단	진단상세내용	<ul style="list-style-type: none"> 데이터 모델과 실제 DB에 구현된 컬럼 단위 불일치 정도를 진단 		
		진단대상	구조 > 컬럼	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 데이터모델에 정의되지 않았으나 실제DB에만 존재하는 경우, 데이터모델에 정의되었으나 실제DB에는 존재하지 않는 경우의 불일치 컬럼(컬럼명, 타입, 길이, 제약조건 등) 		
		오류율 계산	<ul style="list-style-type: none"> 불일치율(%) = (데이터모델기준 누락 컬럼 수 + 실제DB 기준 누락 컬럼 수) / (데이터모델 전체 컬럼수 + 실제DB기준 전체 컬럼수) x 100% 		
	식별자 누락 진단	진단상세내용	<ul style="list-style-type: none"> 데이터를 유일하게 식별 가능한 식별자가 정의되었는지 진단 		
		진단대상	구조 > 테이블	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 전체 테이블 중에서 식별자(Primary Key)가 정의되어 있지 않은 테이블이 있는지 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%) = (식별자 미정의 테이블 수) / (대상DB 전체 테이블 수) x 100% 		
공간객체 완전성	공간객체 누락, 초과 진단	진단상세내용	<ul style="list-style-type: none"> 지적도, 국가기본도 등과 같은 기본공간정보를 기준으로 공간객체를 비교하여 공간객체가 누락되어 있거나 초과하는 공간객체가 존재하는지 진단 		
		진단대상	공간정보	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 기준공간정보와 비교하여 누락된 공간객체 또는 추가로 생성된 공간객체가 없어야 한다. (참조) ArcGis의 topology rule기능을 참조하여 진단 <ul style="list-style-type: none"> 누락 : Must be larger than cluster tolerance 불일치 : Must cover each other 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%) = (오류 폴리곤 수 / 진단 총 폴리곤 수) x 100% 		

1.2 유효성

세부지표	진단내용	분류	진단기준 및 진단방법		
범위 유효성	유효범위 진단	진단상세내용	• 값이 업무 요건에 의해 정해진 유효 범위 내에 존재하는지 진단		
		진단대상	값	진단기법	프로파일링
		진단방법	• 금액, 수량, 윌 도메인 컬럼을 대상으로 업무 요건에서 정의한 유효 범위를 벗어난 데이터가 있는지 진단		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
날짜 유효성	날짜형식 진단	진단상세내용	• 일자 도메인 데이터의 값이 해당 도메인의 일자 형식을 준수하는지 진단		
		진단대상	값	진단기법	프로파일링
		진단방법	• 일자 도메인 데이터의 값이 해당 도메인의 일자 형식을 위반한 데이터가 있는지 진단		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
형식 유효성	패턴 진단	진단상세내용	• 데이터 값이 해당 도메인의 데이터 구성 형식(Pattern)을 준수하는지 진단		
		진단대상	값	진단기법	프로파일링
		진단방법	• 번호 (또는 ID) 도메인의 항목이 정해진 형식을 위반한 데이터가 있는지 진단		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
코드 유효성	코드 유효값 진단	진단상세내용	• 값이 업무 요건에 의해 정해진 해당 코드 도메인의 유효 값 내에 존재하는지 진단		
		진단대상	값	진단기법	프로파일링
		진단방법	• 코드 도메인 컬럼을 대상으로 업무 요건에서 정의한 유효 코드 값에 해당하지 않는 데이터가 있는지 진단		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
여부 유효성	여부 유효값 진단	진단상세내용	• 여부 도메인 데이터의 값이 유효 값 내에 존재하는지 진단		
		진단대상	값	진단기법	프로파일링
		진단방법	• 여부 도메인 컬럼에 유효 코드 값에 해당하지 않는 데이터가 있는지 진단		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		

I

II

III

데이터 품질관리 기준

IV

V

세부지표	진단내용	분류	진단기준 및 진단방법		
선후관계 유효성	업무적 선후관계 유효값 진단	진단상세내용	<ul style="list-style-type: none"> 단일 테이블 내에서 시계열적 선후관계가 업무규칙으로 정의된 경우 발생 업무규칙을 준수하는지 진단 		
		진단대상	값	진단기법	업무규칙
		진단방법	<ul style="list-style-type: none"> 2개 이상의 일자 도메인 컬럼이 업무 규칙에 따른 데이터 발생의 논리적 선후 관계를 위배한 데이터가 있는지 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100% 		
그리드 유효성	위치 정확성 진단	진단상세내용	<ul style="list-style-type: none"> 공간정보의 공간적 위치(격자, 좌표계(타원체/투영체계))가 일치하는 지를 진단하는 지표 		
		진단대상	래스터 공간정보(도엽 : 정사영상, 수치표고모형 등)	진단기법	업무규칙
		진단방법	<ul style="list-style-type: none"> 래스터 파일 해상도와 영상의 경계 값을 확인하여 이미지의 정확한 가로, 세로 픽셀수를 구하고, 계산된 픽셀 수와 이미지 속성정보의 픽셀 수를 비교하여 래스터 파일 해상도와 격자 간격 정확도 진단 		
		오류율 계산	적정 여부		
	범위 유효성 진단	진단상세내용	<ul style="list-style-type: none"> 공간정보의 공간적 위치(격자, 좌표계(타원체/투영체계))가 일치하는 지를 진단하는 지표 		
		진단대상	래스터 공간정보	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 도엽으로 이루어진 래스터(Raster) 공간정보의 공간적 범위가 도엽 및 좌표계의 범위로부터 유효한지 여부를 진단(육안검사) 		
		오류율 계산	적정 여부		
	공간 참조체계 정확성 진단	진단상세내용	<ul style="list-style-type: none"> 공간정보의 공간적 위치(격자, 좌표계(타원체/투영체계))가 일치하는 지를 진단하는 지표 		
		진단대상	공간정보	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> prj 파일에 정의된 좌표계(타원체/투영체계) 및 공간범위의 확인을 통하여 레이어에 속한 객체들의 좌표 (최소값(MIN)과 최대값(MAX))이 주어진 공간적 범위를 벗어났는지를 진단 		
		오류율 계산	적정 여부		

1.3 일관성

세부지표	진단내용	분류	진단기준 및 진단방법		
참조 무결성	참조 무결성 진단	진단상세내용	<ul style="list-style-type: none"> 기본 키와 참조 키 간의 관계가 항상 유지되는지 진단 		
		진단대상	값	진단기법	프로파일링, 업무규칙
		진단방법	<ul style="list-style-type: none"> 외래 키 컬럼이 데이터 값이 참조하는 Parent 테이블에 존재하지 않는 데이터 값이 있는지 진단. 데이터모델(ERD)의 참조 관계가 정의되었으나 실제 DB에는 참조관계 무결성 제약이 정의되지 않은 경우 진단 데이터모델(ERD)에 참조관계 정의가 없지만 업무규칙에 근거하여 참조 관계가 있는 경우는 “업무규칙”을 정의하여 진단 (이러한 경우 데이터모델 보완이 병행되어야 함) 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100% 		
중복 일관성	중복 일관성 진단	진단상세내용	<ul style="list-style-type: none"> 동일한 데이터를 중복하여 관리하는 경우 원천데이터와 중복데이터를 구분할 수 있고, 동일한 시점에 중복데이터가 같은 값을 가지고 있는지 진단 		
		진단대상	값	진단기법	프로파일링, 업무규칙
		진단방법	<ul style="list-style-type: none"> 성능 상의 이유로 중복하여 정의한 컬럼이 데이터 값이 원천 데이터 값과 불일치 데이터가 있는지 진단. (메타데이터에 중복 설계한 컬럼의 원천정보를 관리하지 않는 경우는 업무 담당자가 업무규칙으로 정의하여 진단) 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100% 		
표준 일관성	표준용어 진단	진단상세내용	<ul style="list-style-type: none"> 모델(구조)의 모든 컬럼명이 표준용어를 적용하고 준수하고 있는지 진단 		
		진단대상	표준 >컬럼	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 모델(구조)의 컬럼명 대상으로 표준용어를 준수하지 않은 컬럼이 있는지 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (표준용어 위반 컬럼 수 / 진단 총 컬럼 수) × 100% 		
표준 일관성	표준 도메인 진단	진단상세내용	<ul style="list-style-type: none"> 모델(구조)의 모든 컬럼이 표준도메인을 적용하고 준수하고 있는지 진단 		
		진단대상	표준 >컬럼	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 모델(구조)의 모든 컬럼을 대상으로 표준도메인을 준수하지 않은 컬럼이 있는지 진단 . (컬럼에 적용된 도메인이 표준에 등록되어 있는지, 데이터 타입과 데이터 길이가 표준도메인과 일치하는지 진단) 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (표준도메인 위반 컬럼 수 / 진단 총 컬럼 수) × 100% 		
	엔티티 명명규칙 진단	진단상세내용	<ul style="list-style-type: none"> 모델(구조)의 엔티티명이 명명규칙을 준수하고 있는지 진단 		
		진단대상	표준 >엔티티	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 모델(구조)의 엔티티명이 명명규칙을 준수하지 않은 엔티티가 있는지 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (명명규칙 위반 엔티티 수 / 진단 총 엔티티 수) × 100% 		
	테이블 명명규칙 진단	진단상세내용	<ul style="list-style-type: none"> 모델(구조)의 테이블명이 명명규칙을 준수하고 있는지 진단 		
		진단대상	표준 >테이블	진단기법	프로파일링

I

II

III

데이터 품질관리 기준

IV

V

세부지표	진단내용	분류	진단기준 및 진단방법		
		진단방법	<ul style="list-style-type: none"> 모델(구조)의 테이블명이 명명규칙을 준수하지 않은 테이블이 있는지 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (명명규칙 위반 테이블 수 / 진단 총 테이블 수) × 100% 		
위상구조 일관성	공간객체 위상구조 진단	진단상세내용	<ul style="list-style-type: none"> 공간객체가 기하학적 위상을 준수하여 생성되었는지를 진단 		
		진단대상	공간객체(점, 선, 면)	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 공간개체를 표현하는 점, 선, 면의 위상 표현의 정확성만을 진단 (진단방법) ArcGis의 Check Geometry 기능을 참조하여 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (오류 폴리곤 수 / 진단 총 폴리곤 수) × 100% 		
개념적 일관성	개념적 일관성	진단상세내용	<ul style="list-style-type: none"> 공간객체간의 위상관계 규칙 준수여부를 진단하는 지표 기준공간정보와 생성한 공간정보의 레이어(layer)를 오버랩을 통하여 사전에 정의한 공간금칙규칙에 위배되는 객체 진단 - 철도 레이어에 건물 존재, 호수와 겹쳐지거나 포함된 건물 존재 		
		진단대상	공간객체	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> ArcGis의 기능을 참조하여 진단 Must be larger than cluster tolerance Must not overlap with Must be covered by 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (오류 폴리곤 수 / 진단 총 폴리곤 수) × 100% 		
	개별 공간파일 완결성	진단상세내용	<ul style="list-style-type: none"> 공간객체간의 위상관계 규칙 준수여부를 진단하는 지표 개별 공간정보(레이어) 내의 공간객체들이 공간적 완결성(형식, 관계)을 유지하고 있는지에 대하여 진단 공간정보에 홀 폴리곤(hole polygon) 및 멀티 폴리곤(multi polygon)으로 구성된 공간객체가 영역의 겹침이 없어야 한다. 		
		진단대상	공간객체	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 형식) 개별파일에는 같은 유형의 공간정보로 구성되었는지 진단 관계) 개별 공간파일 내의 객체가 Overlap(겹침)이 있는지 진단 관계) 개별 공간파일 내의 동일 객체가 있는지 진단 ArcGis의 topology rule기능을 참조하여 진단 Must not overlap 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (오류 폴리곤 수 / 진단 총 폴리곤 수) × 100% 		
연계 일관성	연계정보 일치성 진단	진단상세내용	<ul style="list-style-type: none"> 원천 테이블의 연계 대상 컬럼의 데이터와 수집영역의 복제 테이블의 컬럼의 데이터 상호 일치하는지 진단 		
		진단대상	값	진단기법	프로파일링
		진단방법	<ul style="list-style-type: none"> 데이터 송신기관 원천 테이블의 연계 대상 컬럼별 데이터 값과 수신기관 복제 테이블의 컬럼별 수집 데이터 총 건수, 값의 상호 일치 여부를 진단 		
		오류율 계산	<ul style="list-style-type: none"> 오류율(%)= (불일치 컬럼 수 / 진단 총 컬럼 수) × 100% 		

1.4 정확성

세부지표	진단내용	분류	진단기준 및 진단방법		
논리관계 정확성	업무 규정에 따른 데이터 값 진단	진단상세내용	• 업무규정에 근거하여 해당 데이터 값이 정확하게 생성되었는지 진단		
		진단대상	값	진단기법	업무규칙
		진단방법	• 업무규정, 관계법령을 기준으로 업무 관계가 있는 컬럼의 데이터 값이 정확하지 않은 데이터가 있는지 진단		
		오류율 계산	오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
파생항목 정확성	파생 데이터 값 정확성 진단	진단상세내용	• 파생되어 관리되는 항목에 대해 원천 항목의 값과 파생 된 항목의 값이 정확하게 관리되고 있는지 진단		
		진단대상	값	진단기법	업무규칙
		진단방법	• 파생되어 관리되는 컬럼의 값이 원천 항목으로부터 파생규칙에 따라 정확하게 생성되지 않은 데이터가 있는지 진단		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
메타 정확성	비정형 속성정보의 정확성 진단	진단상세내용	• 비정형 콘텐츠의 메타 정보가 해당 비정형 콘텐츠의 속성 정보를 관리 하는 컬럼의 데이터 값과 정확하게 일치하는지 진단		
		진단대상	값	진단기법	업무규칙
		진단방법	<ul style="list-style-type: none"> 비정형 콘텐츠의 메타 정보가 해당 비정형 콘텐츠의 속성 정보를 관리 하는 컬럼의 데이터 값과 다른 데이터가 있는지 진단 (이미지 메타정보(파일포맷, 용량, 압축율, 해상도 등)와 이미지 콘텐츠의 실제 속성정보 불일치 진단) (동영상 메타정보(파일포맷, 용량, 압축률, 화면비율, 초당 프레임수 등)와 동영상 콘텐츠의 실제 속성정보 불일치 진단) (오디오 메타정보(파일포맷, 용량, 압축률, 소리크기 등)와 음성 콘텐츠의 실제 속성정보 불일치 진단) 		
		오류율 계산	• 오류율(%)= 오류 건수(레코드)/ 진단 총 건수(레코드) × 100%		
위치식별 정확성	식별자 정의	진단상세내용	• 기본공간정보를 기준으로 위치 식별자(도로명주소, PNU(지번주소), UFID(공간정보참조체계) 등)가 정확히 구축되어 있는지를 진단		
		진단대상	도로명주소, PNU(지번주소), UFID(공간정보참조체계) 등 속성값 및 공간객체의 위치	진단기법	프로파일링
		진단방법	• 기본공간정보(연속지적도, 도로명주소도 등)를 기준으로 비교하여 진단		
		오류율 계산	• 오류율(%) = 오류 건수(레코드) / 진단 총 건수(레코드) × 100%		

I

II

III

데이터 품질관리 기준

IV

V

02

정성적 진단 기법



데이터 품질지표 정성적 진단 방법은 진단 결과가 YES, NO 등 여부로 도출되는 진단 방법으로 빅데이터 플랫폼 및 센터 품질관리담당자가 지표별 체크리스트를 통해 주관적으로 판단 한다.

값, 구조, 표준, 공간정보, 관리체계 등 모든 진단대상을 진단하며, 품질 지표를 기준으로 품질지표별 진단 특성에 따른 체크리스트로 진단을 수행한다.

표 4 정성적 진단방법 체크리스트

지표	진단특성	체크리스트
완전성	데이터모델	• 데이터 모델 설계 원칙이 수립되어 있다.
	데이터모델	• 데이터 구조 설계 시 논리 데이터 모델을 사용하고 있다.
	데이터모델	• 논리 데이터 모델은 모델링 방법론을 적용하여 충분히 상세한 수준으로 작성되어 있다.
	데이터모델	• 데이터베이스의 구조 변경 시 논리 모델과 물리 모델을 변경한 뒤 데이터베이스에 적용하고 있다.
	공간정보메타(공간정보)	• 공간정보 메타데이터의 각 항목이 누락 없이 기준에 따라 작성하여 관리되고 있다.
유효성	표준	• 데이터 표준(단어, 용어, 도메인, 코드값, 엔티티 및 테이블 명명규칙)이 정의되어 있다.
	관리도구	• 데이터 표준, 데이터모델 정보(엔티티정의서, 속성정의서, 테이블정의서, 컬럼 정의서), DBMS 등은 메타데이터 시스템으로 등록, 변경관리하고 있다.
	관리체계	• 데이터 표준 정의 및 데이터 표준 적용을 위한 절차가 정의되어 있다.
	분류표준	• 공간정보 메타데이터의 분류체계항목이 공간정보표준분류체계(대분류14종, 중분류 45종)를 준용하며 작성되어 관리되고 있다.
	위치표준	• 공간정보 메타데이터의 공간정보기준(축척, 해상도, 측지계, 투영법, 좌표원점, 좌표계)항목이 기준에 따라 작성하여 관리되고 있다.
일관성	표준	• 업무변경 및 표준변경 등에 의한 논리 모델의 변경 및 추가 시 정보시스템에 어떤 영향이 있는지 알 수 있다.
	표준	• 데이터 표준 및 표준화 절차가 정의되어 있으며 이를 준수하고 있다.

지표	진단특성	체크리스트
	표준	• 플랫폼의 표준코드에 등록된 코드를 일관되게 사용하고 있다.
	표준	• 명명규칙에 따른 데이터 표준 준수 현황을 파악하고 있다.
	중복값	• 데이터베이스내의 중복시킨 데이터를 구분하고 정합성을 검증할 수 있다.
	중복값	• 중복시킨 데이터가 정합성을 위배한 경우 이를 보정하기 위한 절차나 방법이 있다.
	연계	• 외부 시스템과 연계하는 데이터의 종류와 연계 주기, 방법을 파악 할 수 있다.
	연계	• 수집 및 제공되는 연계 데이터 관리 절차가 있으며, 이를 준수하고 있다.
	연계	• 수집 및 제공되는 연계 데이터에 대해 데이터 값에 대한 정합성을 검증하고 있다.
	연계	• 외부로부터 수집되는 연계 데이터에 대해 내부 데이터 흐름 및 변경 영향을 파악할 수 있다.
	연계	• 외부에 제공하는 데이터의 변경 발생 시 해당 기관에 통보하는 절차가 정의되어 있다.
	참조관계	• 데이터 간 종속관계가 있거나 참조관계가 있을 경우 데이터베이스에 제약 조건이 설정되어 있거나 프로그램을 통해 검증하고 있다.
	참조관계	• 데이터 간 종속관계 및 참조관계가 있으나 실제 참조되지 않는 데이터가 있을 때 원인을 파악할 수 있다.
	위치표준	• 수평적 연결 관계를 갖은 부분 공간정보들을 통합을 위해 공간정보기준(축척, 해상도, 측지계, 투영법, 좌표원점, 좌표계) 및 해상도를 일관되게 적용하고 있다.
	인코딩표준	• 문자데이터의 인코딩이 국가공간정보센터가 준용하고 있는 문자 인코딩 규정(UTF-8, ASCII)에 따라 메타데이터에 명시하고 있다.
	연계모니터링	• 연계체계가 구축되어 있으며, 연계상태의 모니터링을 수행하고 있다.
정확성	오류처리	• 오류데이터원인을 파악하고 이를 보정하기 위한 절차가 있다.
	업무규칙	• 업무와 관련한 법, 규정, 지침(이하 근거 규정)이 명시된 업무규칙 정의서가 있다.
	업무규칙	• 근거 규정이 변경된 경우 이를 업무규칙 정의서에 반영하기 위한 절차가 있다.
	업무규칙	• 업무규칙 정의서를 기반으로 데이터의 정확성을 검증하고 있다.
	업무규칙	• 원천 데이터를 특정 계산식에 의해 가공하여 활용하는 경우 계산식을 관리하고 있고 이를 통해 데이터 검증을 할 수 있다.
준비성	품질관리체계	• 단기 또는 중·장기 정보화 계획에 데이터 품질관리가 고려되어 있다.
	품질관리체계	• 데이터 품질관리 정책, 지침, 절차가 정의되어 있다.
	품질관리체계	• 데이터 품질관리 업무 수행을 위한 조직 및 역할과 책임이 정해져 있다.
	품질관리체계	• 데이터 품질관리를 위한 표준 문서가 최신의 상태로 유지되고 있다
	품질관리체계	• 데이터 품질관리 정책, 지침, 절차를 준수하고 있다.
	품질관리조직	• 데이터 품질관리를 수행함에 있어 현재 운영 중인 조직은 적정하다
	품질인식제고	• 경영자 및 의사결정권자는 데이터 품질관리의 필요성에 대해 이해하고 있다.
	관리도구	• 데이터 품질을 측정 및 분석하는 도구나 방법이 있으며, 업무 수행에 활용하고 있다.
	관리도구	• 데이터 구조적 품질관리 업무 수행을 위한 데이터모델을 관리하는 도구가 있다.

I

II

III

데이터 품질관리 기준

IV

V

지표	진단특성	체크리스트
보안성	데이터관리책임	• 데이터에 대한 관리 담당자가 지정되어 있다.
	데이터관리책임	• 데이터 소유자에 의해 데이터의 통제(등록, 변경, 삭제)가 이루어지고 있다.
	접근통제	• 데이터 품질관리 표준 문서에 대한 사용자 접근 권한이 설정되어 있다.
	접근통제	• 데이터에 대한 접근 권한이 설정되어 있다.
	접근통제	• 데이터 접근에 대한 사용 이력(Log)을 관리하고 있다.
	정보보호	• 개인정보보호 대상 데이터가 파악되고 접근제한 및 암호화 등을 적용하고 있다.
	정보보호	• 데이터에 관한 보안 감사를 수행하고 발견된 취약점을 개선하고 있다.
	정보보호	• 비공개 대상(저작권, 개인정보 등) 정보를 파악하여 관리하고 있다.
	공간정보공개	• 공간정보 메타데이터의 공개구분 항목이 공개원칙 기준에 따라 작성하여 관리되고 있다.
유용성	충분한 데이터	• 사용자(분석가)의 분석을 위한 충분한 데이터를 제공하고 있다.
	충분한 데이터	• 사용자(활용자)가 활용하기 위한 충분한 데이터를 제공하고 있다.
	활용 편의	• 사용자 편의성을 높이기 위해 데이터를 다양한 형태로 제공하고 있다. (파일형식 : CSV, JSON, XML 등 특정 S/W에 종속되지 않는 오픈 포맷형식으로 제공) (공간정보파일은 shape file 표준 파일포맷으로 제공)
	활용서비스	• 사용자의 요구사항을 수집 및 관리하고 있다.
	활용서비스	• 사용자에게 제공한 데이터는 어떤 용도로 활용되고 있는지 파악할 수 있다.
	활용서비스	• 서비스하고 있는 데이터에 대한 사용자 만족도를 조사하고 있다.
	활용서비스	• 데이터 제공이나 오류처리 등을 위한 처리 절차가 있으며, 준수되고 있다.
	공간정보이력	• 공간정보 메타데이터의 원본데이터소스 항목이 관리되고 있다.
접근성	검색편의	• 사용자가 이용 가능한 데이터를 파악할 수 있도록 정보를 제공하고 있다.
	접근편의	• 사용자가 어떤 경로를 통해 데이터를 접근하고 있는지 파악할 수 있다.
	접근편의	• 데이터 접근 경로를 파악하고 있다면 이를 서비스 개선에 활용하고 있다.
	사용 편의	• 사용자 편의성을 높이기 위해 데이터를 다양한 형태로 제공하고 있다.
	형식표준	• 공간정보(벡터 및 래스터) 파일의 저장형식(확장자)이 국가공간정보센터의 지정형식으로 저장되어 있다.
적시성	적시제공	• 사용자에게 데이터를 제공하기 위한 데이터 제공 주기가 정의되어 있으며 이를 준수하고 있다.
	적시제공	• 사용자의 데이터 처리 요청은 처리기한 내에 반영되고 있다.
	최신데이터	• 실제 업무에서 변경된 값을 DB에 반영하기 위한 규칙과 절차가 정의되어 있다.
	최신데이터	• 데이터의 갱신 주기와 방법이 문서화 되어 있고 이를 적용하고 있다.
	정보갱신	• 공간정보 메타데이터의 정보생성 및 갱신 항목이 관리기준에 따라 작성되어 있으며, 데이터의 최신성을 유지하고 있다.

01

개요



빅데이터 플랫폼 및 센터 데이터 품질관리 기준을 정립하고 데이터 관리 및 연계·활용 효과와 효율성 향상을 위해 데이터 품질관리에 관한 수준 평가체계 마련이 필요하다.

데이터 품질관리가 체계적으로 이루어지도록 수준을 측정할 수 있는 성숙도 평가 모델 체계 마련이 필요하며, 데이터 품질 관리 성숙도 평가 모델은 크게 평가체계 정립, 평가기준정의, 평가방법을 정의하고 이에 따라 데이터 성숙도 평가 항목을 도출하여 단위 조직 또는 전사 측면에서 성숙도 평가를 수행하도록 데이터 성숙도 평가 모델을 수립한다.

▶ 성숙도 평가체계 정립

- 국내·외 성숙도 평가 모델 대상 분석하고 주요 구성요소를 파악하여 데이터 품질 관리 성숙도 평가체계 수립의 기초자료로 활용한다.
- 성숙도 평가 모델의 구성요소는 데이터 품질기준(지표), 데이터 품질관리 요소(프로세스), 데이터 품질관리 프로세스 성숙 수준 등 3가지 구성요소를 기본으로 데이터 품질관리 성숙도 평가 모델 프레임워크를 구축한다.

▶ 성숙도 평가기준 정의

- 데이터 품질관리 성숙도 평가기준은 9개 데이터 품질지표와 품질지표를 향상시키기 위해서 필요하거나 품질지표에 영향을 주는 프로세스를 식별하여 12가지 프로세스를 도출 하였고 프로세스와 품질지표와의 연관 관계에 따라 상호 매핑한다.

▶ 성숙도 평가방법 설계

- 데이터 품질관리 성숙도 평가모델은 크게 데이터 값(Value) 진단을 실시하여 그 결과 데이터의 오류율(%)을 기준으로 평가하는 「데이터 값 수준평가」와 데이터 품질관리 프로세스별 체크리스트 진단 후 각 프로세스별 관리 수준을 평가하는 「데이터 관리체계 수준평가」로 데이터 품질관리 성숙도 평가방법을 설계한다.

▶ 성숙도 평가항목 도출

- 데이터 품질관리 성숙도 평가항목은 12개 프로세스별로 성숙도 수준 평가를 최하위 수준 1에서 최상위 수준 5로 평가항목별 관리기준, 평가항목, 증빙자료, 충족 요건 등을 제시하여 점진적으로 성숙 수준이 향상될 수 있도록 평가 항목을 정의한다.

I

II

III

데이터 품질관리 기준

IV

V

02

평가 체계



2.1 평가체계 동향분석

데이터 성숙도 평가체계는 기존 국내·외 성숙도 모델을 분석하고 각 모델의 주요한 특징과 구성요소를 파악하여 본 사업에 적용 가능한 성숙도 평가체계를 정의한다.

- 국내 성숙도 모델 분석 대상 선정 : 데이터 품질지표와 데이터 품질관리 성숙도 모형(DQM3: Data Quality Maturity Model 3), EA(Enterprise Architecture)성숙도 모형, 녹색정보화 성숙도 모형, 공공데이터 품질관리 수준평가 모델을 분석
- 국외 성숙도 모델 분석 대상 선정 : SW조직의 프로세스 기반 능력과 성숙도 수준평가 모델의 CMMI(Capability Maturity Model Integrate), SW 프로세스의 지속적인 개선을 위한 평가 모델의 ISO 15504 (SPICE: Software Process Improvement and Capability dEtermination), 데이터 품질을 통한 비즈니스 성과향상의 DQAT(Data Quality Assessment Tool), IT거버넌스 차원의 광범위한 프로세스 수준평가의 COBIT(Control OBjectives for Information and related Technology) 모델을 분석
- 국내·외 성숙도 모델의 주요 특징과 구성요소 파악 : 국내·외 성숙도 모델의 주요 특징은 데이터 및 프로세스 기반의 성숙 수준을 평가하고, 품질, 프로세스, 성숙도를 평가하는 모델로 데이터 품질 공인 인증 서비스를 지원하는 평가 모델이다.

표 5 국내외 성숙도 평가 모델 분석 현황

구분	분석 대상 모델	주요 특징	구성 체계		
			품질지표	평가등급	프로세스
국내	DQM3	• 품질지표와 프로세스 기반의 성숙도 평가	6개	5단계	8개
	EA성숙도 모델	• 중앙행정기관과 광역자치단체 구분하여 평가	-	5단계	9개
	녹색정보화 수준	• 녹색정보화 수준진단 및 지표별 성숙도 평가	64개	6단계	15개
	공공데이터 수준	• 기관 자체평가 결과를 바탕으로 확인점검 실시	4개영역	5단계	9개
해외	CMMI-Dev	• SW조직 프로세스 기반 능력&성숙도 수준 평가	-	5단계	22개
	ISO 15504	• SW 프로세스의 지속적인 개선을 위한 평가 모델	-	5단계	48개
	DQAT	• 개선된 데이터 품질을 통한 비즈니스 성과향상	-	5단계	-
	COBIT	• IT거버넌스 차원 광범위한 프로세스 수준 평가	-	6단계	32개

2.2 평가체계 프레임워크

2.2.1 평가체계 구성요소

데이터 성숙도 평가체계 수립을 위한 성숙도 평가 구성요소는 데이터 품질기준(지표), 데이터 품질관리 요소(프로세스), 데이터 품질관리 프로세스 성숙 수준의 3가지를 기본으로 한다.

첫 번째 축인 데이터 품질기준(지표)은 데이터 품질의 다양한 측면을 분류·정리해 완전성, 유효성, 일관성, 정확성, 준비성, 보안성, 유용성, 접근성, 적시성의 9가지 품질 지표로 정의한다.

두 번째 축인 데이터 품질관리 요소(프로세스)는 품질기준(지표)을 향상시키기 위해서 필요하거나 품질기준(지표)에 영향을 주는 프로세스를 식별하여 12가지 프로세스를 도출 했다.

프로세스 참조 모델의 Plan-Do-Check-Act Cycle 형태를 기반으로 설계한 프로세스는 품질관리계획, 데이터 품질관리 역량, 데이터 품질기준관리(DQI, CTQ, BR), 데이터 오너십 관리, 데이터 표준 관리, 데이터 구조관리, 데이터 연계관리, 데이터 보호관리, 품질진단 및 개선관리, 오류 데이터 관리, 요구사항관리, 사용자 뷰 관리가 포함된다.

세 번째 축인 데이터 품질관리 프로세스 성숙 수준은 조직(플랫폼 및 센터)의 데이터 품질관리 발전 단계를 고려해 ‘초기화-관리-정의-정량적 관리-최적화’의 5수준 체계로 구분 하고, 12개의 프로세스 별로 성숙도 수준 평가를 통해 최종 달성수준을 부여하도록 정의 한다.

즉, 『데이터 품질관리 성숙도 수준 평가체계』는 데이터 품질기준(지표)과 관련 프로세스별 연관 관계 매핑을 통해 각 프로세스별 성숙 수준 평가 항목을 측정하고, 프로세스별 달성 정도를 평가 후 해당 기관의 최종 성숙도 수준 평가는 최하위 프로세스 평가결과를 기준으로 부여토록 설계한다.

2.2.2 성숙도 수준 평가 추진 절차

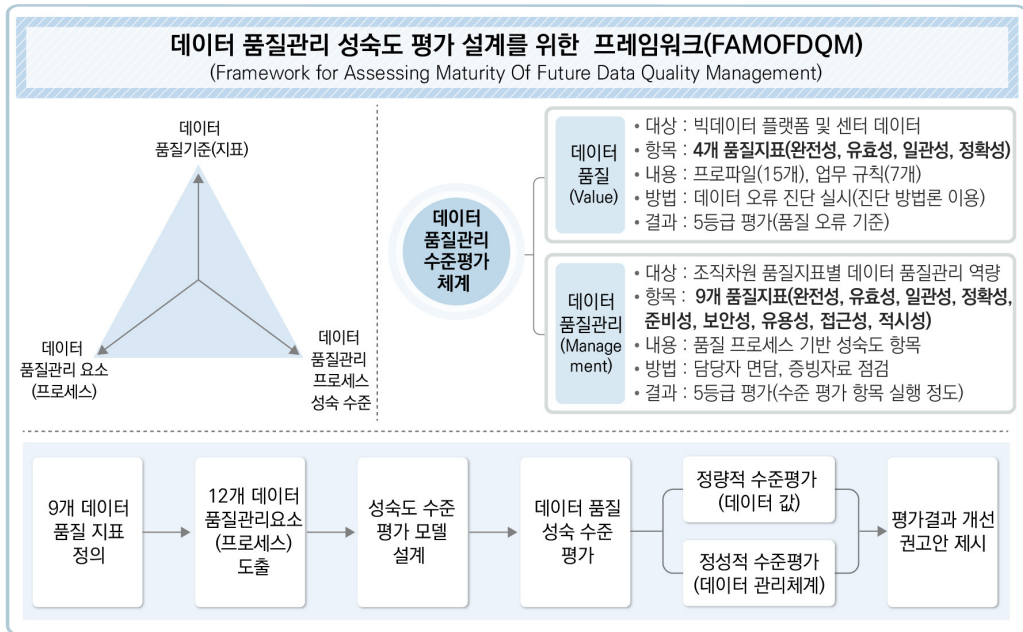
데이터 성숙도 수준 평가 추진 절차는 성숙도 품질기준(9개 지표) 정의, 데이터 품질관리 요소(12개 프로세스), 데이터 품질관리 프로세스 성숙 수준의 3가지 구성요소를 기반으로 성숙도 수준 평가 모델을 설계한다. 이를 바탕으로 데이터 성숙도 수준 평가는 데이터 값(Value)의 정략적 수준평가와 데이터 관리체계(Management)의 정성적 수준 평가를 실시한 후, 평가 결과에 대해 해당 프로세스 별로 개선 권고안을 제시하는 절차에 따라 수행한다.

2.2.3 성숙도 수준 평가 모델 프레임워크(Framework)

『데이터 품질관리 성숙도 수준 평가 모델 프레임워크』는 데이터 품질관리 3가지 구성요소(데이터 품질기준(지표), 데이터 품질관리 요소(프로세스), 데이터 품질관리 프로세스 성숙 수준)를 기반으로 크게 데이터 값(Value)을 측정하는 부분과 데이터 품질관리 체계 (Management)를 측정하고 평가하는 체계로 구성한다.

이를 체계적으로 평가하기 위한 성숙도 수준 평가 절차에 따라 성숙도 수준 평가를 실시하기 위한 성숙도 수준평가 모델 프레임워크를 수립한다.

그림 6 데이터 품질관리 성숙도 수준 평가 모델 프레임워크



03

평가기준



데이터 품질관리 성숙도 수준 평가는 데이터 품질관리 성숙도 수준 평가 모델 프레임워크에서 제시했던 바와 같이, 데이터 값에 대한 진단 및 평가와 데이터 품질관리 체계에 대한 평가를 체계적이고 객관적으로 측정한다. 데이터 품질관리 활동의 전반적인 사항을 포함하여 최종 성숙도 수준을 평가하기 위한 기준을 제시한다.

3.1 성숙도 평가 기준 정의

데이터 품질관리 성숙도 수준 평가는 데이터 품질관리 요소(12개 프로세스)를 기반으로 각 프로세스별 평가항목의 달성 여부를 평가(Y/N)하고 각 수준별 평가 항목을 모두 충족했을 때 해당 수준을 충족한 것으로 평가한다.

3.1.1 각 단계별 프로세스 정의

데이터 품질관리 프로세스 참조 모델인 PDCA Cycle 모델을 기반으로 데이터 생명주기를 고려하여 [계획-구축-운영-활용] 단계로 구성하고, 각 단계 프로세스별 성숙도 수준 평가 항목의 충족 여부(Y/N))를 평가하여 5단계(수준 1~수준 5)로 평가한다.

표 6 단계별 각 프로세스 정의

단계	프로세스	설명
계획	데이터 품질관리 계획	• 데이터 품질관리 계획 수립, 표준 프로세스 수립, 배포, 품질관리 이행 점검, 중장기 계획 수립 등
	데이터 품질관리 역량	• 데이터 품질관리 조직 역량 강화를 위한 교육훈련 계획, 실행, 노하우 지식 축적 및 공유
	데이터 품질기준관리 (DQI, CTQ, BR)	• 대상후보 도출, AP영향분석, 업무규칙 생성, 핵심 품질항목(CTQ) 도출, 데이터 품질지표(DQI) 정의
	데이터 오너십 관리	• 데이터 조직, 역할 및 책임정의, 담당자 지정
	데이터 표준관리	• 데이터 표준(코드, 단어, 용어, 도메인)계획 수립, 점검 및 조치, 개선 활동 수행

단계	프로세스	설명
구축	데이터 구조관리	• 논리/물리 데이터모델 정의, 점검 및 조치, 구조관리 최적화
	데이터 연계관리	• 연계 메타데이터 정의, 변경관리, 정합성 진단, • 흐름 및 추적관리, 연계 데이터 품질개선
	데이터 보호관리	• 정보보호 관리계획, 보안감사, 취약점 점검 계획, 진단, 결과분석, 대응 및 조치
운영	데이터 품질진단 및 개선관리	• 진단정의, 진단 수행 및 개선, 근본원인 분석 및 해결, 품질통제
	오류 데이터 관리	• 데이터 값 진단을 통해 데이터 오류율 평가
활용	요구사항 관리	• 요구사항 요청 접수, 요구사항 결과분석, 활용
	사용자 뷰 관리	• 사용자의 데이터 활용 극대화, 데이터 이용 환경 관리

3.1.2 성숙도 수준 달성 정도

각 단계 프로세스의 성숙도 수준 평가는 각 수준별 평가 항목의 달성 여부(Y/N)를 평가하고, 모든 평가 항목을 만족할 때 해당 수준의 성숙도를 달성했다고 평가한다.

각 프로세스 성숙도 평가 항목은 수준별 평가 항목을 정의하고, 달성 정도를 객관적으로 확인하기 위해 해당 평가 항목의 이행 결과에 대한 증빙자료와 충족 기준을 제시한다.

3.2 성숙도 수준 평가 체계

데이터 품질관리 성숙도 수준 평가 체계는 최하위 수준 1에서부터 최상위 수준 5까지 5단계 수준 평가 체계로 정의한다.

일반적으로 조직의 성숙도 수준 평가는 능력(Capability) 수준과 성숙도(Maturity) 수준으로 분류하여 평가를 실시하고 있음을 고려하여 본 사업의 성숙도 수준 평가는 이를 조합하여 단계별 프로세스별 해당 조직(플랫폼 및 센터) 데이터 품질관리 프로세스 전반적인 달성 정도(실행 상태)를 평가하여 프로세스 성숙도 달성 수준이 가장 낮은 프로세스를 기준으로 최종 성숙도 수준 평가를 판정한다.

일부 성숙도 수준 평가 모델에서는 ‘등급’으로 평가하여 최하위 등급을 5등급으로 하고 최상위 등급을 1등급으로 표현하는 경우도 있으나 본 사업에서는 CMMI 등 선진 성숙도 수준 평가 체계를 인용하여 최하위 수준 1에서 최상위 수준 5로 성숙도 수준 평가 체계를 정의한다.

3.3.1 수준 1 (초기화)

데이터 품질관리가 인식이 미흡하여 기본적인 품질관리 활동의 수행이 불가능 하거나 부분적인 품질관리 활동만 수행되는 수준이다.

기본적인 데이터 품질관리 활동이 시작된다. 수준1에서는 데이터 오류에 대한 적발 및 단순한 데이터 오류 정제 등 사후 조치가 가능하며 특정 DB나 단위 조직의 단기적인 품질 향상에 집중한다.

3.3.2 수준 2 (관리)

데이터 품질관리가 인식되고 품질진단에 따른 개선 조치 등 기본적인 품질관리 활동 들을 수행하는 수준이다.

프로세스 중심의 상시적인 품질관리 기반이 구축된다.

수준2에서는 데이터 값에 영향을 주는 표준관리, 구조관리, 연계관리, 사용자 오너십 관리 등의 통제 범위가 확대되며 데이터 오류에 대한 예방중심으로 활동이 강화된다. 수준1과 수준2의 차이는 일정한 형태로 수행에 들어가면 수준1이 시작되었다고 간주하며 수행을 하되 정해진 절차를 따라서 수행하면 수준2가 시작된 것이다.

3.3.3 수준 3 (정의)

데이터 품질관리를 위한 전반적인 활동이 관리 및 통제되어 이를 통해 데이터 품질 향상이 가능한 수준이다.

단위 조직 중심의 데이터 품질관리가 조직 전체 차원으로 확대된다.

수준 3에서는 조직 전체 차원의 데이터 품질관리를 효과적으로 지원하기 위해 데이터의 성능, 장애, 보호, 백업 등의 데이터 운영관리가 필요하며, 조직 차원의 반복적 품질 이슈에 관한 원인분석 및 개선방안 수립이 가능하다.

수준 2와 수준 3의 차이는 수준 3은 이미 정형화 되었다는 의미가 있으며, 수행에 필요한 규정과 절차가 준비되어야 하며 이러한 절차나 규정 및 프로세스 등이 마련되어 있지만 실제 상황에서는 다소 지켜지지 않는다면 수준 2로 간주한다.

3.3.4 수준 4 (정량적 관리)

데이터 품질관리 활동이 조직 전체적으로 프로세스가 이행되고, 데이터 품질관리 활동 수행에 따른 성과측정이 가능한 수준이다.

수준 3의 원인분석 및 개선방안 제시에 따른 프로세스 개선이 수행된다. 수준 4에서는 프로세스 개선에 따른 성과측정 기준을 마련하여 개선 전·후의 점검이 가능해져 프로세스의 개선 효과가 검증된다.

3.3.5 수준 5 (최적화)

데이터 품질관리 활동이 전사적인 관점에서 데이터 품질관리 활동의 선순환 체계가 확립되고, 이를 통해 빅데이터 플랫폼 및 센터의 데이터 품질관리가 안정적 품질 향상 및 유지가 보장되는 수준이다. 또한, 조직 전체의 지속적인 프로세스 개선이 가능하고, 이를 통해 조직의 데이터 품질관리에 대한 단계적 목표 달성이 가능하다.

수준 5인 최적화 상태에서는 조직에 맞는 품질 성과관리 체계가 확보되고, 지속적으로 평가되어 목표 수립이나 계획 수립에 반영되는 선순환 체계가 구축된다.

표 7 성숙도 수준 체계

수준	표현	정의
수준 5	최적화	<ul style="list-style-type: none"> 조직 전체의 데이터 품질관리 활동의 선순환 체계가 확립 안정적 품질 향상 및 유지 보장
수준 4	정량적 관리	<ul style="list-style-type: none"> 조직 차원의 데이터 품질관리 프로세스 이행 데이터 품질관리 활동이 체계적으로 수행
수준 3	정의	<ul style="list-style-type: none"> 데이터 품질관리 전반적인 활동이 관리 및 통제 이를 통해 데이터 품질 향상이 가능
수준 2	관리	<ul style="list-style-type: none"> 데이터 품질관리 인식 품질진단 등 기본적인 품질관리 활동들을 도입·시작
수준 1	초기화	<ul style="list-style-type: none"> 데이터 품질관리 인식이 미흡 기본적인 품질관리 활동 불가능하거나 부분적인 품질관리 활동 수행

현재 성숙도 수준 평가에서 상위 성숙도 수준 달성을 위해서는 하위 성숙 수준을 모두 달성해야 함
(예 : 현재 성숙도 수준이 수준 3이라면 수준 1~2의 성숙 수준이 완전 달성(Fully) 되고, 수준 3의 성숙 수준이 달성(Fully or Largely) 되어야 함.

3.3 최종 성숙도 수준 평가

해당 기관(조직)의 데이터 품질관리 성숙도 수준에 대한 최종 평가는 각 프로세스 평가항목을 측정 (Y/N)한 후 해당 수준별 평가항목이 모두 충족되었을 때 해당 수준의 성숙도를 달성했다고 평가한다. 따라서 해당 기관의 데이터 품질관리 성숙도 수준 최종 평가는 최하위 수준의 프로세스 성숙도 수준 평가 결과를 기준으로 판정한다.

04

평가방법



데이터 품질관리 성숙도 수준 평가방법은 데이터 품질관리 성숙도 수준 평가 모델 프레임워크에서 정의한 바와 같이 첫 번째는 데이터 품질기준(지표)을 기반으로 데이터 값(Value)에 대한 진단을 통해 품질지표별 오류율(%)에 대한 수준 평가를 실시한다.

두 번째는 데이터 품질관리 프로세스를 기반으로 데이터 관리체계(Management)를 수준별로 도출된 성숙도 평가 항목을 점검하여 평가점수를 수준 달성정도의 판정 기준에 따라 최종 수준을 부여한다.

4.1 성숙도 수준 평가 : 데이터 값(Value)

데이터 값(Value)에 대한 성숙도 수준 평가는 데이터 진단 대상을 선정하고, 데이터 품질 진단 대상에 대해 정량적 진단 기법인 프로파일링과 업무규칙에 의해 정량적 품질지표 (완전성, 유효성, 일관성, 정확성)별 진단항목을 진단한다.

그리고 데이터 값에 대한 오류 범위를 6시그마 관점에서 정하고, 데이터 정합률(%)을 기준으로 성숙도 수준을 정의한다.

4.1.1 프로파일링 진단

프로파일링 진단 대상인 품질지표별로 품질진단을 실시한다.

▶ 완전성 품질지표 진단

구조 완전성의 테이블 중복진단, 테이블 현행화 진단, 컬럼 현행화 진단, 식별자 누락 진단 내용을 필수 값과 관계 값에 대해 진단을 실시한다.

▶ 유효성 품질지표 진단

코드 유효성의 코드 유효값, 형식 유효성의 패턴진단, 여부 유효성의 여부 유효값, 날짜 유효성의 날짜형식에 대해 여부, 수량, 금액, 월, 날짜, 코드 진단 항목에 대해 진단을 실시한다.

범위 유효성의 금액, 월, 수량에 대해 유효범위를 진단한다.

▶ 일관성 품질지표 진단

표준 일관성의 표준용어 진단, 표준 도메인 진단, 엔티티 명명규칙 진단, 테이블 명명 규칙에 대해 시간순서, 컬럼논리관계, 계산 및 집계 등의 진단항목을 진단한다.

▶ 정확성 품질지표 진단

메타 정확성의 비정형 속성정보의 정확성에 대해 번호, 명칭 등의 진단항목을 진단한다.

4.1.2 업무규칙 진단

업무규칙 진단 대상인 품질지표별로 품질진단을 실시한다.

▶ 완전성 품질지표 진단

단독 완전성의 Null 값, 조건 완전성의 Null 값에 대해 진단한다.

▶ 유효성 품질지표 진단

선후관계 유효성의 업무적 선후관계 유효성에 대해 진단한다.

▶ 일관성 품질지표 진단

참조 무결성에 대해 진단한다.

▶ 정확성 품질지표 진단

논리관계 정확성의 업무규정에 따른 데이터 값에 대해 진단하고, 파생항목 정확성의 파생 데이터 값 정확성에 대해 진단한다.

표 8 데이터 값 품질진단

프로파일링			업무규칙		
품질지표		진단내용	품질지표		진단내용
완전성	구조 완전성	• 테이블 종목 진단	완전성	단독완전성	• Null 값 진단
		• 테이블 현행화 진단			
		• 컬럼 현행화 진단		조건완전성	• Null 값 진단
		• 식별자 누락 진단			
	공간객체 완전성	• 공간객체 누락, 초과 진단			
유효성	코드 유효성	• 코드 유효값	유효성	선후관계	• 업무적 선후관계 유효값 진단
	형식 유효성	• 패턴 진단			
	여부 유효성	• 여부 유효값 진단			
	날짜 유효성	• 날짜형식 진단			
	범위 유효성	• 금액, 울, 수량 유효범위			
	그리드 유효성	• 위치 정확성 진단			
		• 범위 유효성 진단			
		• 공간참조체계 정확성 진단			
일관성	표준 일관성	• 표준 용어진단	일관성	참조 무결성	• 참조 무결성 진단
		• 표준 도메인 진단			

프로파일링			업무규칙		
품질지표		진단내용	품질지표		진단내용
		• 엔티티 명명 규칙 진단	중복 일관성	• 중복 일관성 진단	
		• 테이블 명명 규칙 진단			
	위상구조 일관성	• 공간객체 위상구조 진단			
	개념적 일관성	• 개념적(스키마) 일관성			
		• 개별 공간 파일 완결성			
정확성	메타 정확성	• 비정형 속성정보 정확성 진단	정확성	논리관계	• 업무규칙에 따른 데이터 값
	위치식별 정확성	• 식별자 진단		파생항목	• 파생 데이터 값 정확성 진단

4.1.3 데이터 값 수준 평가 판정

정량적 품질진단 대상인 완전성, 유효성, 일관성, 정확성 4개의 품질지표별로 진단항목을 진단하고, 품질지표별 전체 건수 대비 오류 건수의 오류율(%)을 산정한다.

표 9 데이터 값 진단 항목 및 진단결과 오류율(%) 산정(예시)

품질지표별 진단항목		진단 결과(예시)		
품질지표	진단 항목	전체 건수	오류건수	오류율(%)
완전성	필수 값, 관계 값	100	2	2.00%
유효성	여부, 수량, 금액, 율, 날짜, 코드	235	6	2.55%
일관성	시간순서, 컬럼 논리관계, 계산 및 집계	500	21	4.20%
정확성	번호, 명칭	350	16	4.57%
계		1,185	45	3.80%

4.1.4 데이터 값 성숙도 수준 평가

데이터 값에 대한 성숙도 수준 평가는 등급별 성숙도 평가를 지양하고 6시그마의 높은 수준이 달성되는 수준이어야 의미 있는 데이터 품질을 확보함을 보장함으로 데이터 값 성숙도 수준 평가는 6시그마 관점에서 데이터 정합률(%)을 평가한다.

표 10 데이터 값 성숙도 수준 평가

구 분		데이터 수준 평가
Ace class	4.9σ이상	데이터 품질 정합률 99.97% 이상
High class	3.8σ이상	데이터 품질 정합률 99.0% 이상
Middle class	3.7σ이상	데이터 품질 정합률 98.5% 이상
Low class	3.7σ미만	데이터 품질 정합률 98.5% 미만

4.1.5 데이터 값 성숙도 수준 평가 절차

데이터 값에 대한 성숙도 수준평가 절차는 데이터 진단대상을 선정하고, 정량적 진단의 프로파일링과 업무규칙에 대한 품질지표별 진단 후 전체 건수 대비 오류건수에 대한 오류율(%)을 산정하고, 6시그마 관점에서 데이터 정합률(%)에 의해 성숙도 수준 평가를 한다.

4.2 성숙도 수준 평가 : 데이터 관리체계(Management)

데이터 관리체계(Management)에 대한 성숙도 수준 평가 체계는 9개 품질지표와 관련 프로세스와의 상호 매핑을 통해 12개 프로세스별 성숙도 평가 항목을 도출한다.

이를 기반으로 해당 조직의 프로세스별 수준 평가를 실시 한 후 '3. 평가기준'에서 정의한 바에 따라 데이터 관리체계에 대한 수준 평가(1~5수준)를 한다.

4.2.1 프로세스별 수준 평가 항목 도출

데이터 품질관리 수준 평가는 9개 품질지표와 관련 프로세스와의 상호 매핑을 통해 데이터 품질관리 성숙도 수준 평가를 위해 프로세스별 성숙도 수준 평가 항목을 도출한다.

평가 항목을 기반으로 해당 조직(플랫폼 및 센터)의 데이터 품질관리에 대한 성숙도 수준 평가 활동을 실시한다. 프로세스별 성숙도 수준 평가 항목 구성에 대한 구체적인 사항은 '5. 평가항목'에서 정의한다.

4.2.2 성숙도 수준 평가

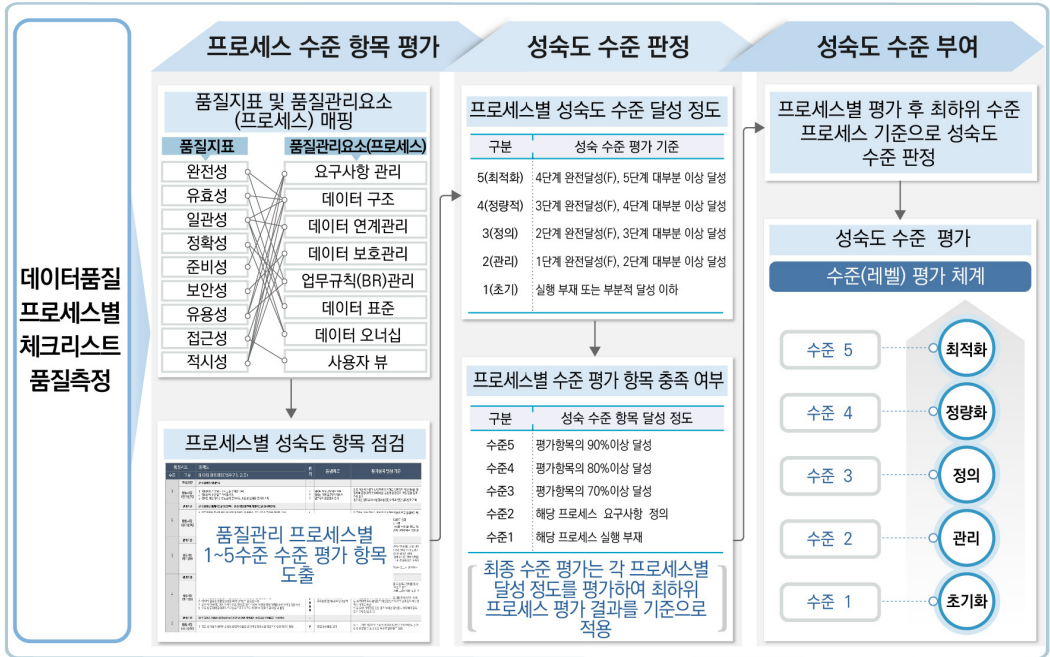
각 단계 프로세스별 성숙도 평가 항목을 기반으로 각 프로세스별 성숙도 수준의 달성 정도를 평가(Y/N)하고, 각 프로세스별 수준 평가항목의 달성정도가 모두 충족했을 때 해당 프로세스 성숙도 수준이 달성되었다고 평가한다.

4.2.3 데이터 관리체계 성숙도 수준 평가

데이터 관리체계 성숙도 수준 평가는 각 프로세스별 성숙도 수준 평가 항목이 모두 충족했을 때 해당 프로세스의 성숙도 수준이 달성되었다고 평가하고, 해당 기관(조직)의 데이터 관리체계 성숙도 수준 최종 평가는 각 프로세스 성숙도 수준 평가 결과 최하위 수준 평가 프로세스를 기준으로 최종 평가를 판정한다.

데이터 품질관리 성숙도 수준 평가 체계는 '3.2 성숙도 수준 평가 체계'에서 정의한 바에 따라 최하위 수준(수준 1)에서 최상위 수준(수준 5)의 평가를 각 프로세스별로 평가하고, 최하위 평가 프로세스를 해당 기관(조직)의 최종 성숙도 수준 평가로 판정한다.

그림 7 데이터 품질관리 성숙도 수준 평가 절차



I

II

III

데이터 품질관리 기준

IV

V

05

평가항목



데이터 품질관리 성숙도 수준 평가를 위한 프로세스별 평가항목은 해당 기관(조직)별로 데이터 품질 관리 수준 평가 활동을 수행하면서 적용하는 체크리스트 형식으로 프로세스별 성숙도 수준 평가(1~5수준) 체계로 항목을 정의 한다.

5.1 프로세스별 평가항목 정의

데이터 품질관리 수준 평가 항목은 각 단계(계획, 구축, 운영, 활용)의 12개 프로세스로 평가항목을 도출하고, 최하위 수준 1에서 최상위 수준 5의 순서로 평가항목을 측정하도록 정의하였다.

5.2 평가항목 구성 체계

각 단계 정의, 관련 프로세스, 수준 1 항목은 수준 2 평가 항목을 측정하여 미 정립 또는 부분적 수행의 경우 수준 1 부여, 성숙도 수준 평가항목(관리수준, 평가항목, 충족기준), 평가(충족, 미 충족), 증빙자료(평가항목의 충족, 미충족의 판단 근거)로 구성한다.

5.2.1 평가항목 구성요소

데이터 품질관리 성숙도 수준 평가항목은 표 4-11 성숙도 평가 항목에서 정의한다.

표 11 성숙도 평가 항목

구성항목	설 명	비 고
단계	• 계획, 구축, 운영, 활용의 4단계로 구성	
관련 프로세스	• 12개 프로세스 명 정의(예 : 데이터 표준관리 프로세스)	• 단계별로 정의
수준	• 단계 프로세스별 평가 수준(수준 5~수준 2)	• 수준 1 항목 없음
관리기준	• 해당 평가항목 수행에 대해 기본적으로 관리하는 기준 정의	
평가항목	• 수준별 개별 평가항목의 요구사항(질의사항)	
충족기준	• 평가항목을 평가 시 기본적으로 충족하는지 여부 측정 기준	• 모두 충족
평가	• 평가항목의 측정 시 수준별 충족여부 판정 • “충족” : 평가항목에 부합하게 증빙자료 제공 및 확인 • “미 충족” : 평가항목 조건을 충족하지 못하고 증빙자료 미 제공	• 해당 프로세스별 평가 항목의 수준 평가를 모두 충족
증빙자료	• 프로세스별 평가항목의 활동에 대한 객관적 근거의 증빙자료	• 문서, 시스템

5.3 단계 프로세스별 평가항목

5.3.1 계획단계 프로세스 - 데이터 품질관리 계획 프로세스 수준 평가 항목

단계		계획	평가	증빙자료
관련프로세스		데이터 품질관리 계획		
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 데이터 품질관리계획 수립		
	평가항목	1. 데이터 품질관리계획을 수립하고 있습니까? - 데이터 품질관리 계획을 수립하고 공식화하고 있음		• 데이터품질관리 계획서
		2. 데이터 품질관리 업무 수행을 위한 조직과 역할 및 책임이 명확하게 정의하고 있습니까? - 데이터 품질관리 조직, 담당자, 역할과 책임(R&R)을 정의하고 있음		• 조직도, 업무분장 내역서(역할&책임)
		3. 데이터 품질관리 기본문서들은 주기적으로 현행화하여 관리하기 위한 절차를 정의하고 있습니까? - 데이터 품질관리 기본문서의 주기적 현행화 하는 절차가 상세하게 정의, 절차에 따라 현행화 활동을 수행하고 있음		• 데이터 품질관리 현황 분석서
	(충족기준)	<ul style="list-style-type: none"> 데이터 품질관리 계획이 승인권자의 승인을 득하고 공식화하여야 함 데이터 품질관리 계획이 별도의 계획으로 작성되어야 하는 것은 아니며, 다른 상위 계획에 포함된 경우도 인정 		
3	관리기준	3-1 데이터 품질관리 정책, 지침, 절차를 준수(이행점검 및 조치)(계속)		
	평가항목	1. 데이터 품질관리계획 대비 추진실적 점검 및 조치(70% 이상) 활동을 수행합니까? - 데이터 품질관리계획 대비 추진실적을 점검하고, 점점에 따른 조치활동(조치계획, 조치실적 등)을 수행하였음 - 품질관리계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치활동을 70%이상을 수행 - 측정산식 : · 계획 대비 실적 달성률(%) = (추진계획 수행 건수 / 데이터 품질관리 계획의 추진과제 계획 건수)*100		• 데이터품질관리 추진실적점검 결과서
		2. 데이터 품질을 측정 및 분석하는 도구나 방법이 있으며, 업무 수행에 활용합니까? - 데이터 품질관리 업무 수행을 위한 품질관리 도구를 구축하고 활용함		• 품질관리도구 활용
		3. 데이터 품질관리 기본문서에 대한 접근 권한을 설정하고 이를 준수합니까?		• 접근권한설정 및 이행 결과서

I

II

III

데이터 품질관리 기준

IV

V

단계	계획		평가	증빙자료
관련프로세스	데이터 품질관리 계획			
수준	성숙도 수준 평가 항목			
		- 데이터 품질관리 기본문서의 접근권한 설정 및 이행 준수해야 함		
	(충족기준)	<ul style="list-style-type: none">데이터 품질관리 계획서에 포함된 추진과제(산출물 관리 계획 등)에 대한 수행여부의 점검 결과를 제시해야 함계획 대비 실적 달성률은 70% 이상 달성되었음을 보장해야 함		
4	관리기준	4-1 데이터 품질관리 이행점검 및 조치(계속)		
	평가항목	1. 데이터 품질관리 계획 대비 추진실적 점검 및 조치(80%이상) 활동을 수행합니까? - 데이터 품질관리 계획 대비 추진실적을 점검하고, 점검에 따른 조치활동(조치계획, 조치실적 등)을 수행하였음 - 데이터 품질관리 계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치활동을 80% 이상 수행		<ul style="list-style-type: none">데이터 품질관리 추진실적 점검결과서
		2. 품질 향상을 위해 성과관리 활동을 수행합니까? - 지속적인 품질향상을 위한 품질개선 활동과 성과관리 활동을 수행하고 있음		<ul style="list-style-type: none">성과 관리 결과서
		3. 점검 및 조치결과를 보고 및 공유하고 있습니까? - 점검 및 조치결과를 의사결정권자에게 보고하고, 관련 이해관계자와 공유(배포)하고 있음		<ul style="list-style-type: none">데이터 품질관리 추진실적 점검결과서
	(충족기준)	<ul style="list-style-type: none">데이터 품질관리 계획 대비 실적에 포함하여 점검하고 조치 수행결과를 확인하여 수행되고 있음공식화된 종합 결과보고서는 물론 공식화 되지 않은 실제 시스템 구축 화면, 표준화 산출물 등의 개선조치 사실을 확인할 수 있는 증적은 인정		
5	관리기준	5-1 데이터 품질관리 이행점검 및 조치(계속)		
	평가항목	1. 데이터 품질관리 계획 대비 추진실적 점검 및 조치(90% 이상) 활동을 수행 합니까? - 데이터 품질관리 계획 대비 추진 실적을 점검하고, 점검에 따른 조치활동(조치계획, 조치실적 등)을 수행하였음(계속) - 데이터 품질관리 계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치활동을 90% 이상을 수행		<ul style="list-style-type: none">데이터 품질관리 추진실적 점검결과서
	(충족기준)	<ul style="list-style-type: none">당해 연도 중장기 데이터 품질관리 계획을 수립하고 첫 해로써 당해 과제를 수행하는 경우도 인정중장기 데이터 품질관리 계획을 수립하고 개선과제의 이행실태를 점검하고 개선조치 활동을 수행하고 있음		

5.3.2 계획단계 프로세스 - 데이터 품질관리 역량 프로세스 수준 평가 항목

단계	계획	평가	증빙자료
관련프로세스	데이터 품질관리역량		
수준	성숙도 수준 평가 항목		
2	관리기준 2-1 데이터 품질관리 교육계획 수립		
	평가항목 <ul style="list-style-type: none"> 1. 데이터 품질관리 교육계획을 수립하고 있습니까? <ul style="list-style-type: none"> - 해당 조직의 데이터 품질관리 역량강화를 위한 연간 품질관리 교육훈련 계획을 수립하여 공식화하고 있음 - 품질관리 교육훈련 계획에는 품질관리 교육목표, 교육대상, 주요 교육과정 내용, 교육일정 등의 내용을 포함하고 있어야 함 		<ul style="list-style-type: none"> • 데이터 품질관리 교육훈련 계획서
	(충족기준)		<ul style="list-style-type: none"> • 데이터 품질관리 교육훈련 계획은 반드시 별도의 계획으로 작성되어야 하는 것은 아니며, 다른 상위 계획에 포함된 경우도 인정
	관리기준 2-2 데이터 품질관리 담당자 지정		
3	평가항목 <ul style="list-style-type: none"> 1. 데이터 품질관리 담당자 지정하고 있습니까? <ul style="list-style-type: none"> - 해당 조직은 데이터 품질관리를 위해 품질관리 담당자를 지정하여 관리하고 있음 - 조직 품질관리 총괄 담당자가 지정되어야 함 - 해당 조직의 업무분장표 등에 역할이 명시되어 있어야 함 		<ul style="list-style-type: none"> • 데이터 품질관리 담당자 지정내역 (업무분장표 등)
	(충족기준)		<ul style="list-style-type: none"> • 해당 조직의 업무분장표 등에 명시되어 있어야 하며 해당 조직 홈페이지, 내부 게시판, 품질관리계획 등에 기재되어 공유되어야 함
3	관리기준 3-1 데이터 품질관리 담당자 지정(계속)		
	평가항목 <ul style="list-style-type: none"> 1. 데이터 품질관리 담당자 지정(수준평가 대상 DB 70% 이상) 하고 있습니까? <ul style="list-style-type: none"> - 정보시스템 또는 DB별 품질관리 담당자(1명 이상)를 지정하여 관리하고 있음 - 해당 조직의 담당자를 지정해야 하며, 수준평가 대상 DB의 70%이상이 DB품질관리 담당자를 지정하고 있어야 함 - 측정산식 : $\text{데이터 품질관리 담당자 지정률(\%)} = (\text{담당자가 지정된 DB 개수} / \text{수준평가 대상 DB개수}) * 100$ 		<ul style="list-style-type: none"> • 데이터 품질관리 담당자 지정내역
	관리기준 3-2 데이터 품질관리 교육 실시		
	평가항목 <ul style="list-style-type: none"> 1. 데이터 품질관리 교육훈련을 실시(연간 1회)하고 있습니까? <ul style="list-style-type: none"> - 데이터 품질관리 계획에 따라 해당 조직 자체적으로 DB 품질관리 담당자 및 이해관계자(정보화 담당자, 업무 담당자)를 대상으로 데이터 품질관리 교육을 실시하고 있음(연간 1회 이상) - 데이터 품질관리 교육은 회당 2시간 이상의 교육시간을 충족하고 있으며, 연간 기준으로 품질관리 담당자는 1회 이상 교육에 참여하여야 함 - 해당 조직차원의 품질관리 대상 DB가 존재하지 않은 경우 '데이터 품질관리 담당자 지정' 관련 항목 해당없음(N/A)으로 처리하여 평가 - 데이터 품질 담당자는 기본적으로 해당 조직 소속 담당자여야 하나, 공공기관의 경우 부처의 소속기관이나 산하기관의 담당자로 지정된 경우도 인정 - 범정부 차원에서 실시하는 품질관리 교육에 참여한 경우도 해당조직 자체 교육 실시로 인정 		<ul style="list-style-type: none"> • 데이터 품질관리 교육자료 • 데이터 품질관리 교육 결과서 및 보고 내역 • 교육 참석 확인할 수 있는 문서 (교육참석 확인증, 관련 공문 등)

I

II

III

데이터 품질관리 기준

IV

V

단계 관련프로세스 수준	계획 데이터 품질관리역량 성숙도 수준 평가 항목	평가	증빙자료
	- 품질관리 실무 교육 인정 범위는 범정부 차원의 교육을 모두 인정(품질관리 기본, 실무, 전문 교육 등)		
(충족기준)	<ul style="list-style-type: none">교육 대상자는 수준평가 진단 대상 DB의 품질관리 담당자임교육 대상자의 전원이 교육에 참여해야 함교육 횟수와 시간에 상관없이 조직의 총 시간을 충족하면 인정(2시간)해당 조직 자체평가 중 혹은 그 이후 담당자가 변경되었을 경우 업무변경 사실이 확인되면 담당자가 교육 참여한 것으로 인정		
관리기준	4-1 데이터 품질관리 담당자 지정(계속)		
평가항목	1. 데이터 품질관리 담당자 지정(수준평가 대상 DB 80%이상)을 하고 있습니까? - 데이터 품질관리 담당자를 지정(1명 이상)하여 관리하고 있음 - 해당 조직 담당자를 지정해야 하며, 수준평가 대상 DB의 80%이상 DB에 대하여 품질관리 담당자가 지정되어야 함		<ul style="list-style-type: none">데이터 품질관리 담당자 지정내역
관리기준	4-2 데이터 품질관리 교육훈련 실시(계속)		
4	평가항목	1. 데이터 품질관리 교육훈련 실시(연간 2회 이상)하고 있습니까? - 데이터 품질관리 계획에 따라 해당 조직 자체적으로 DB 품질관리 담당자 및 이해관계자(정보화 담당자, 업무 담당자)를 대상으로 데이터 품질관리 교육을 실시하고 있음 - 데이터 품질관리 교육은 회당 3시간 이상의 교육 시간을 충족하고 있으며 연간 기준으로 2회 이상 교육에 참여해야 함 - 교육훈련 계획에 따라 실시된 교육 결과에 대한 성과 평가를 분석하여 상위관리자에게 결과보고를 수행하고 있음	<ul style="list-style-type: none">데이터 품질관리 교육자료데이터 품질관리 교육 결과서 및 보고 내역교육참석 확인할 수 있는 문서(교육참석 확인증, 관련 공문 등)
(충족기준)	<ul style="list-style-type: none">교육은 횟수와 시간에 상관없이 해당 조직의 총 시간을 충족하면 인정(6시간)		
관리기준	5-1 데이터 품질관리 담당자 지정(계속)		
평가항목	1. 데이터 품질관리 담당자 지정(수준평가 대상 DB 90%이상)을 하고 있습니까? - 데이터 품질관리 담당자를 지정(1명 이상)하여 관리하고 있음 - 해당 조직 담당자를 지정해야 하며, 수준평가 대상 DB의 90%이상 DB에 대하여 품질관리 담당자가 지정되어야 함		<ul style="list-style-type: none">데이터 품질관리 담당자 지정내역
관리기준	5-2 데이터 품질관리 노하우 경험의 축적 및 확산		
5	평가항목	1. 데이터 품질관리 이해관계자 간 소통을 정례화(연간 2회 이상)하고 있습니까? - 품질관리 담당자 및 이해관계자(정보화 담당자, 업무 담당자 등)가 참여하는 실무차원의 데이터 품질관리 의사소통 활동을 실시하고 있어야 함(연 2회 이상) - 데이터 품질관리를 주제로 세미나, 토론회, 간담회 등을 통해 회당 1시간 이상, 연간기준으로 2회 이상 참여하여야 함 - 의사소통 등은 교육 종료 후 연속해서 진행되는 경우도 인정	<ul style="list-style-type: none">데이터 품질관리 의 사소통 회의자료데이터 품질관리 관 련 노하우, 경험, 교 육, 사례, 지식 등이 문서화된 자료
(충족기준)	<ul style="list-style-type: none">의사소통 활동은 참여자 모집 공지 및 의견 기록물 등의 구체적인 증적을 제출해야 함의사소통의 내용(예시)<ul style="list-style-type: none">품질관리 담당자 및 이해관계자가 참여하는 의사소통 활동의 주기적인 수행의사결정권자(조직의 장, 책임자 등)에게 품질관련 주여 사항 정기적인 보고		

5.3.3 계획단계 프로세스 - 데이터 품질기준 관리(업무규칙)프로세스 수준 평가 항목

단계	계획			
관련프로세스	데이터 품질기준 관리(DQI, CTQ, 업무규칙(BR))		평가	증빙자료
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 관리대상을 체계적으로 정의, 계획적으로 수행		
	평가항목	1. 업무 규칙의 특성에 따라 체계적인 업무규칙을 정의하고 산출물을 관리합니까? - 컬럼의 필수 여부 및 유효 범위, 컬럼 간의 관계 등 유형별 도출 방법을 정의하고 있음 - 중복컬럼에 대한 데이터 검증 규칙을 정의하고 관리 - 참조무결성 검증을 위한 규칙을 정의하고 관리 - 데이터 품질 진단의 우선순위 및 중요도관리를 위한 핵심품질 항목(CTQ)을 정의하고 관리		<ul style="list-style-type: none"> 업무규칙 정의서
	(충족기준)	<ul style="list-style-type: none"> 업무규칙 대상정의서(대상 테이블, 대상 컬럼, 중요도(테이블, 컬럼), 영향도(서비스, 업무))를 작성하여 관리해야 함 중복 데이터 관리 목록(예 : 시스템명, Primary(주제영역, 엔티티명, 테이블명, 속성명, 컬럼명), Child(주제영역, 엔티티명, 테이블명, 속성명, 컬럼명))을 정의하고 관리해야 함 핵심품질항목 우선순위 선정목록(예 : 정보그룹, 핵심품질항목, 업무중요도, 업무영향도, 데이터 집중도, 결함도, 등 점수화하여 순위, 관리단계 등)을 정의하고 관리해야 함 		
	관리기준	2-2 데이터 오류에 대한 조치 규칙 및 방법 정의		
	평가항목	1. 데이터 오류에 대한 조치 규칙 및 방법을 정의하고 산출물을 관리합니까? - 오류 원인을 분석하고 관련 사항에 대한 개선조치 방안을 수립 함 - 오류 원인분석, 개선조치 회의록 및 합의한 사항을 근거문서로 작성하여 해당 업무 규칙과 연계하여 저장, 관리하고 있음		<ul style="list-style-type: none"> 오류조치 계획서
	(충족기준)	<ul style="list-style-type: none"> 오류 원인을 분석하고 관련사항(업무규칙, 관련 테이블명, 컬럼명, 오류발생주기, 오류데이터 유형, 오류 원인분석, 개선 및 조치사항, 정제 관련 사항 등)을 작성하여 저장 및 관리해야 함 		
	관리기준	2-3 업무 규칙에 대한 검증을 위한 측정 방법 및 절차를 정의한다.		
	평가항목	1. 업무 규칙에 대한 대상과 측정방법 및 절차를 수립하고 수행 결과에 대한 산출물을 관리합니까? - 업무규칙 도출 절차 1) 데이터 현상분석 : 데이터 품질 관점의 주요 데이터 항목 식별 및 분석 방법 선택 2) 업무규칙 도출 : 업무규칙 도출 대상 선정, -업무규칙 정의성 작성 -업무규칙 도출 결과 종합 및 이슈 정리 3) 데이터 품질 측정 : -품질점검 환경 준비, -품질 측정		<ul style="list-style-type: none"> 업무규칙정의서 (업무규칙측정방법, 결과 등) 업무규칙별 오류율 측정결과서
	(충족기준)	<ul style="list-style-type: none"> 업무규칙 도출 절차(데이터 현상분석, 업무 규칙 도출, 데이터 품질측정)를 정의하고, 업무규칙에 따른 데이터 측정 규칙 수행결과와 업무 규칙별 오류율 측정결과를 관리해야 함 - 업무규칙별 오류율 측정 결과(예시 : BR-ID, 업무, 테이블명, 컬럼명, 품질지표, BR명, 총 		

I

II

III

데이터 품질관리 기준

IV

V

단계	계획			
관련프로세스	데이터 품질기준 관리(DQI, CTQ, 업무규칙(BR))		평가	증빙자료
수준	성숙도 수준 평가 항목			
	건수, 오류건수, 오류율(%)를 관리해야 함			
3	관리기준	3-1 업무규칙에 대한 정확성 검증을 주기적으로 수행		
	평가항목	1. 업무 규칙의 중요도, 오류 도출 정도, 시스템 자원 등을 고려한 검증 주기를 계획하여 수행하십니까? - 데이터 검증 프로그램이나 데이터 품질관리 자동화 도구를 도입하여 주기적으로 수행하여 검증 결과를 상시적으로 모니터링 함 - 검증 결과에 대한 모니터링 결과를 주기적으로 관리자에게 보고하는 체계를 정립하고 증빙자료를 관리해야 함		• 변경영향도 평가서
	(총족기준)	• 업무규칙 정의서의 변경항목(시스템명, DB명, 테이블명, 컬럼명, 변경구분, 변경사항, 변경사유, 우선순위, 중요도 등을 포함)에 대해 데이터 검증 프로그램이나 자동화 도구를 활용하여 주기적으로 검증해야 함		
	관리기준	3-2 전사 시스템 간의 연계 데이터에 대한 체계적인 오류 추적을 수행		
	평가항목	1. 데이터 오류에 대한 조치 규칙 및 방법이 존재하고 이를 적용하니까? - 연계 데이터에 대한 정확성 검증이 가능한 업무규칙을 관리해야 함 - 오류 데이터 개선을 위한 업무규칙 담당자를 정의해야 함 - 오류와 연관된 데이터베이스 또는 어플리케이션에 대한 원인 및 담당자 추적이 가능하고 오류에 대한 대응 활동을 추적하고 기록하여 관리해야 함		• 데이터오류 조치 규칙 및 방법 • 업무분장 내역서
	(총족기준)	• 주요 오류데이터 발생 원인을 데이터 입력, 데이터 흐름, 데이터 재검증 활동별로 구분하여 데이터 품질 개선 절차에 따라 전사적으로 오류추적이 가능해야 함 - 데이터 입력 오류발생원인(이용자 입력 오류, 입력통제 미비, 입력절차 문제점, 입력 프로그램 오류) - 데이터 흐름 오류발생원인(데이터 변환 오류, 데이터 정제 오류, 데이터 적재 오류, 운영자 일괄 변경오류 등) - 데이터 재검증 활동 오류발생 원인(데이터 재검증 절차 및 관리 프로세스 미흡 등)		
	관리기준	3-3 데이터 업무규칙 품질점검 및 개선활동 수행		
	평가항목	1. 데이터 업무규칙에 대한 품질점검(1회/분기 이상)을 실시하니까? - 전사적으로 활용하고 적용하는 데이터 오류에 대한 품질점검(오류 대응 활동)을 실시하고 개선사항 발굴 및 조치하고 있음 - 업무 규칙별 오류율 측정 결과 확인(총 건수 대비 오류 건수)		• 데이터오류 조치 규칙 및 방법 • 업무규칙별 오류율 측정결과
	(총족기준)	• 업무 규칙별 오류율을 측정하고 오류 원인별로 개선 수행 내역을 포함해야 함		
4	관리기준	4-1 데이터 업무규칙 품질 개선활동 수행		
	평가항목	1. 데이터 업무규칙에 대한 품질 개선활동을 수행하니까? - 조치 가능한 데이터 업무 규칙 품질 개선활동 수행		• 데이터 오류 조치 규칙 및

단계		계획		평가	증빙자료
관련프로세스		데이터 품질기준 관리(DQI, CTQ,업무규칙(BR))			
수준	성숙도 수준 평가 항목				
		<div>- 개선사항 발굴 결과, 개선 내역 등을 승인권자에게 보고하고 관련 이해관계자와 공유하고 있음</div> <div>- 데이터 업무규칙 담당자 및 이해관계자가 참여하여 업무규칙 품질 향상을 위한 정기적인 의사소통 활동을 수행하고 있음(1회/분기)</div>			<div>방법</div> <div><div>• 업무규칙별 오류율측정 결과서</div></div>
	(충족기준)	<div>• 데이터 업무규칙 품질개선 활동에 대한 공식화된 증거를 제출해야 함</div> <div>• 의사소통 활동은 4/4분기 각 1회의 활동을 실시하는 것이 원칙이나 연간 합산하여 2회를 충족해도 인정</div>			
	관리기준	4-2 데이터 오류 관리 지식을 축적하여 활용			
	평가항목	<div>1. 오류 활동 내역을 체계적으로 분류하고 지식으로 축적하여 향후 오류 대응 시 이를 적극적으로 활용합니까?</div> <div>- 오류 유형 사례집 등 진단결과에 대한 정보를 사례화하여 공유 문서로서 저장, 관리해야 함</div>			<div>• 오류유형 사례집</div>
	(충족기준)	<div>• 오류 유형을 정의하고 오류사례와 개선방향 등을 지식으로 축적하여 교육, 세미나 등을 통해 공유해야 함</div>			
5	관리기준	5-1 업무규칙을 적용한 품질개선 활동이 직접적인 경영개선 활동으로 연계되어 수행			
	평가항목	<div>1. 업무 규칙을 적용하여 수행한 품질개선 활동 결과가 경영개선을 위한 의사결정 자료로 활용합니까?</div> <div>- 중장기 개선 계획에 따른 데이터 업무규칙 품질개선 활동을 수행하고 있음</div> <div>- 품질개선 활동 결과가 비즈니스 측면에서 경영 개선을 위한 의사결정 자료로 활용됨을 보장함</div>			<div>• 품질개선 활동결과와 경영개선 의사결정 자료 활용 근거자료</div>
	(충족기준)	<div>• 비즈니스 규모가 확대될수록 비즈니스를 수행하는 조직의 규모와 비즈니스를 위해 비즈니스를 위해 실행되는 프로세스가 다양해지고 복잡해짐</div> <div>• 업무 규칙을 적용하여 수집된 품질 측정 자료는 비즈니스 환경을 보완할 수 있는 경영 개선 도구로 활용됨을 보장해야 함</div>			
	관리기준	5-2 오류의 근본 원인 제거 작업 수행			
	평가항목	<div>1. 오류의 근본원인을 파악하여 이를 해결하기 위한 계획을 수립하고 관련 업무 프로세스를 개선합니까?</div> <div>- 오류의 발생 원인을 파악하여 근본원인을 제거하고 해결하는 개선활동을 수행해야 함</div>			<div>• 오류근본원인 제거작업 (근본원인파악, 해결 위한 계획, 관련 프로세스 개선)</div>
	(충족기준)	<div>• 오류의 근본원인을 파악, 이를 해결하기 위한 계획 수립과 관련 업무 프로세스 개선 활동이 수행됨을 보장</div>			

I

II

III

데이터 품질관리 기준

IV

V

5.3.4 계획단계 프로세스 - 데이터 오너십 관리 프로세스 수준 평가 항목

단계	계획		평가	증빙자료
관련프로세스	데이터 오너십 관리			
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 데이터에 대한 오너십 활동 기준 정의		
	평가항목	1. 주제영역, 업무영역 레벨에서 데이터 오너십 부여기준을 정의하고 있습니까? - 데이터 오너십 부서(담당자)를 정의하여 관리해야 함 - 데이터 오너십 부서(담당자)의 활동 기준을 정의해야 함		• 오너소관부서별 역할과 책임 내역서
	(충족기준)	• 주제영역 단위의 오너십을 부여하기 위해서 단위조직별로 테이블 단위의 오너십을 부여하고 활동을 수행해야 함 - 오너 소관 부서별 책임과 역할 정의 (부서, 역할과 책임, 비고 등) (주제영역, 설명, 테이블, 오너부서, 관련부서) • 활동 기준(예 : IT부서는 오류 데이터 추출 및 정비 데이터 이력관리와 기술적 문제에 해당하는 오너십 부여, 현업은 주관부서와 보조부서로 구분하여 역할과 책임을 부여해야 함)		
3	관리기준	3-1 전사적으로 오너십을 부여하고, 책임과 역할에 대한 활동을 수행		
	평가항목	1. 연계데이터에 대한 오너십 부여 및 책임과 역할을 정의하고 있습니까? - 데이터 오너십 관리가(IT파트의 편의성이 아닌) 전사 차원에서 공유 하고 있음 - 전사 데이터에 대해 데이터 오류 개선 시, 담당자의 참여(승인, 회의참여, 개선작업 등) 활동을 기록 관리해야 함		• 연계데이터 담당자 지정 (역할) 내역서 • 전사차원의문제 해결및공유 • 데이터변경이력 추적결과서
	(충족기준)	• 조직 내에서는 부서 간 또는 타 조직 간에 제공되는 정보에 대한 품질향상이 중요해서 데이터 오너십이 명확하게 부여되어야 함 • 데이터 오류 개선활동에 대한 담당자 참여 활동이 기록으로 확인되어야 함		
4	관리기준	4-1 데이터 오너십 관련 품질지표를 관리하고 이를 향상하기 위한 활동을 전사적으로 수행		
	평가항목	1. 데이터 품질현황에 대한 현재 수준을 수치로 관리합니까? - 데이터 개선 사항에 대한 내역을 정량적 성과로 측정해야함		• 사전 진단 결과 수치화, 공표성과 측정결과서
	(충족기준)	• 현재 수준을 파악하기 위해 사전 진단을 통해 진단 결과를 수치화하여 공표해야 함 • 지속적인 추진에 의한 개선 결과를 정기적(1회/분기)으로 측정하고 성과에 대한 모니터링 해야 함 (예: 데이터 오너십 관련 성과지표 기준 : 모니터링 대상(BR도출, 데이터 정제), 해당 데이터, 성과지표(BR도출 달성률, 건수), 평가 대상(데이터 오너))		
5	관리기준	5-1 데이터 오너십 관련 품질개선 활동이 직접적인 경영개선 활동으로 연계되어 수행		
	평가항목	1. 데이터 오너십 관리의 품질개선 활동 결과가 경영개선을 위한 의사결정 자료로 활용합니까? - 데이터 오너십 관리의 품질개선 활동 결과가 경영개선을 위한 활동 자료로 제공됨을 보장해야 함		• 품질개선 활동 결과서
	(충족기준)	- 데이터 오너십 관리를 통해 수집된 품질 측정 자료는 비즈니스 환경을 보완할 수 있는 경영 도구로 활용됨을 제시해야 함		

5.3.5 계획단계 프로세스 - 데이터 표준관리 프로세스 수준 평가 항목

단계	계획			
관련프로세스	데이터 표준관리			
수준	성숙도 수준 평가 항목		평가	증빙자료
2	관리기준	2-1 데이터 표준관리 계획 수립 및 표준을 정의하고 유지		
	평가항목	1. 데이터 표준관리 계획을 수립합니까? - 기관 내 일관되게 적용해야 할 데이터 표준(표준코드, 표준단어, 표준용어, 표준도메인)의 관리를 위한 추진계획을 수립하고 공식화(승인 획득)하고 있음		<ul style="list-style-type: none"> 데이터 표준관리 계획서(표준코드, 표준단어, 표준용어, 표준 도메인)
	(총족기준)	<ul style="list-style-type: none"> 표준관리 계획에는 표준화 적용 대상, 표준 적용 확산 방안, 표준 준수 여부 점검 및 조치 등의 계획이 포함되어야 함 조직에서 사용하고 있는 데이터를 분석하여 향후 공통으로 사용 할 표준 코드, 표준단어, 표준용어, 표준 도메인에 대한 기준을 수립하고, 수립된 기준에 실제 사용한 표준 코드 사전, 표준 도메인 사전을 생성하고 승인을 득한 후 배포해야 함 예)표준 도메인 사전관리(도메인명, 도메인정의, 데이터 타입, 데이터 길이, 소수점 이하 길이, 도메인 분류 등)를 생성하고 배포해야함 		
	관리기준	2-2 데이터 표준관리를 위한 원칙, 가이드라인 등을 지침화 하고 변경 프로세스를 정의 및 관리		
	평가항목	1. 데이터 표준관리를 위한 원칙, 가이드라인 등을 지침화하여 문서로 관리합니까? - 조직에서 사용하는 도메인과 코드를 유형화 하는 지침을 문서화하고 관리해야 함		<ul style="list-style-type: none"> 데이터 표준화 지침서(표준코드, 표준단어, 표준용어, 표준도메인준수운영절차 등 포함)
		2. 데이터 표준의 변경 프로세스를 관리합니까? - 표준 도메인 준수를 위한 운영절차를 정의하고 관리해야 함		
	(총족기준)	<ul style="list-style-type: none"> 조직에서 사용하는 도메인과 코드를 유형화하는 지침을 문서화하여 관리해야 함 (데이터 표준화 지침서 : 데이터 표준화 개요, 표준 단어, 표준용어, 표준 도메인, 데이터 오브젝트 명명규칙, 표준화 운영 조직 및 프로세스 등 포함) 표준 도메인 준수 운영 절차, 표준코드 변경관리 절차가 문서화로 정의되어야 함 		
	관리기준	2-3 데이터 표준(표준코드, 표준단어, 표준용어, 표준 도메인) 정의		
	평가항목	1. 데이터 표준을 정의하고 있습니까? - 데이터 표준 코드를 기관 표준코드를 포함하여 정의해야 함 - 표준용어를 정의하고 이를 준수하기 위한 절차가 존재함 - 정의된 표준코드는 주요 관리항목('코드명', '코드 값', '코드 값 의미')을 포함하고 있어야 함 - 정의된 표준용어는 주요관리항목(용어명, 영문명, 영문약어명, 용어정의)을 포함하고 있어야 함(표준용어에는 도메인을 포함하여 정의되어야 함) - 데이터 표준 도메인을 기관 표준 도메인을 포함하여 정의하고 있음 - 정의된 표준 도메인은 주요관리항목(도메인명, 도메인정의, 데이터 타입, 데이터 길이, 도메인 분류)을 포함하고 있어야 함		<ul style="list-style-type: none"> 표준코드 정의서 표준단어정의서 표준용어정의서 표준도메인정의서
		2. 표준용어에 대한 최신성을 유지하고 있습니까? - 표준용어의 변경이 발생할 경우 현행화를 통해 최신성을 유지해야 함		<ul style="list-style-type: none"> 데이터 표준화 지침서(표준코드, 표준단어, 표준용어, 표준도메인현행화)
		3. 표준데이터 관리절차가 존재하고 준용하고 있습니까?		

I

II

III

데이터 품질관리 기준

IV

V

단계	계획			
관련프로세스	데이터 표준관리		평가	증빙자료
수준	성숙도 수준 평가 항목			
	- 표준용어 사전 관리 절차, 모델 적용 절차 등 표준데이터 관리 절차가 존재하고 이를 준용하여 데이터 모델을 생성하고 변경해야 함			
	(충족기준)	<ul style="list-style-type: none"> 표준코드정의서는 코드명, 코드 값 의미 항목이 반드시 포함되어야 함 표준용어정의서는 용어명, 영문명, 영문약어명, 용어정의, 도메인 항목이 포함되어야 함 		
3	관리기준	3-1 데이터 표준관리 체계 이행점검 및 조치 수행		
	평가항목	1. 데이터 표준 관리 계획 대비 점검 및 조치(70% 이상) 활동을 수행합니까? - 데이터 표준관리 계획 대비 추진 실적을 점검하고, 점검에 따른 조치활동을 수행하였음 - 표준관리 계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치 활동을 70% 이상 수행함 • 계획 대비 실적 달성률(%) = (추진계획 수행건수/표준관리계획의 추진과제 계획 건수) *100		<ul style="list-style-type: none"> 데이터 표준관리 추진실적 점검 결과서
	(충족기준)	<ul style="list-style-type: none"> 추진실적은 계획서에 계획된 과제에 대한 점검 및 조치가 포함되어야 함 		
	관리기준	3-2 전사적으로 표준화를 통하여 일관성을 유지 전사적으로 데이터 표준에 근거하여 데이터 모델 생성		
	평가항목	1. 데이터 표준 현황분석을 합니까? - 기관 표준과 DB표준 간의 데이터 표준을 비교분석, 중복제거 등 표준 현황 파악 및 정비 활동 등을 수행해야 함 2. 데이터 표준 적용을 70% 이상 수행합니까? - 정의된 데이터 표준이 평가 대상 DB에 적용하여 활용되고 있음 - 데이터 표준(용어+도메인) 적용률이 평가대상 컬럼의 70%이상 적용되어야 함 - 측정산식 : · 데이터 표준 적용률(%) =(평가 대상 DB의 데이터 표준 (용어+도메인)이 적용된 컬럼 수/전체 컬럼 수)*100 (1) 전체컬럼수 : 평가 대상 DB의 전체 컬럼 수(중복 제거된 개수) (2) 평가대상 DB의 컬럼(용어+도메인이 모두)이 적용된 컬럼 수 : 실 DB테이블 컬럼(용어+도메인 비교컬럼) 기준 중복 제거된 개수 3. 데이터 표준 변경이력관리를 합니까? - 데이터 표준(코드, 단어, 용어, 도메인)의 생성, 변경, 삭제 이력을 기록하여 관리하고 있음		<ul style="list-style-type: none"> 데이터 표준 현황 분석서 데이터 표준관리 추진실적 점검 결과서 데이터 표준 (코드, 단어, 용어, 도메인) 각각의 변경 내역서(시스템 포함)
	(충족기준)	<ul style="list-style-type: none"> 데이터 표준 적용률이 평가 대상 컬럼의 70% 이상 적용되어야 함 		
4	관리기준	4-1 데이터 표준관리 체계 이행점검 및 조치를 수행		
	평가항목	1. 데이터 표준 관리 계획 대비 점검 및 조치(80% 이상)활동을 수행합니까? - 데이터 표준관리 계획 대비 추진 실적을 점검하고, 점검에 따른 조치활동을 수행하였음		<ul style="list-style-type: none"> 데이터 표준관리 추진실적 점검 결과서

단계	계획			
관련프로세스	데이터 표준관리		평가	증빙자료
수준	성숙도 수준 평가 항목			
		- 표준관리 계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치 활동을 80% 이상 수행하였음		
	(총족기준)	• 데이터 표준관리 계획 대비 점검 및 조치가 80% 이상 실행했음을 증적자료로 제시해야 함		
	관리기준	4-2 데이터 표준 관련 품질지표를 관리하고 이를 향상하기 위한 활동을 전사적으로 수행		
		1. 데이터 표준 적용을 80% 이상 수행합니까? - 정의된 데이터 표준이 평가 대상 DB에 적용하여 활용되고 있음 - 데이터 표준(용어+도메인) 적용률이 평가대상 컬럼의 80% 이상 적용해야 함		• 데이터 표준 현황 분석서
	(총족기준)	• 데이터 표준 매핑은 기관 표준 전체 목록과 DB표준 전체 목록을 포함하여야 함 • 기관 표준과 DB표준의 교집합, 차집합 확인이 가능해야 함 *(교집합) 기관표준과 매핑된 DB표준, (차집합)기관 표준에 매핑되지 않은 DB표준, (차집합) DB표준에 매핑되지 않은 기관 표준		
5	관리기준	5-1 데이터 표준관리 체계 이행점검 및 조치를 수행		
	평가항목	1. 데이터 표준 관리 계획 대비 점검 및 조치(90% 이상) 활동을 수행합니까? - 데이터 표준관리 계획 대비 추진 실적을 점검하고, 점검에 따른 조치활동을 수행하였음 - 표준관리 계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치 활동을 90% 이상 수행		• 데이터 표준관리 추진실적 점검 결과서
	(총족기준)	• 데이터 표준관리 점검 및 조치가 체계적으로 수행되고 해당 프로세스 최적화 상태를 보장함		
	관리기준	5-2 데이터 표준 관련 품질지표를 관리하고 이를 향상하기 위한 활동을 전사적으로 수행		
	평가항목	1. 기관 표준 DB를 반영을 합니까? - 기관 표준을 DB(원천 DB, 개방대상 데이터 파일 등)에 반영하는 활동을 수행하고 있음 2. 데이터 표준 적용을 90% 이상 수행합니까? - 정의된 데이터 표준이 평가 대상 DB에 적용하여 활용되고 있음 - 데이터 표준(용어+도메인) 적용률이 평가대상 컬럼의 90% 이상 적용되어야 함 - 공공의 경우 범정부 차원의 데이터 표준을 평가대상 DB에 반영하기 위한 활동을 수행하고 있음		• 데이터표준 현황 분석서
	(총족기준)	• 공공의 경우 범정부 표준을 반영하기 위한 활동에는 DB에 표준을 적용하거나, 적용 가능 여부를 판단하기 위한 활동 등이 있음		
	관리기준	5-3 데이터 표준 개선 활동이 직접적인 경영개선 활동으로 연계되어 수행		
	평가항목	1. 데이터 표준관리의 품질개선 활동 결과가 경영개선을 위한 의사결정 자료로 활용합니까? - 데이터 표준관리의 품질개선 활동 결과가 직접적으로 경영개선을 위한 자료로 활용하고 있음		• 품질개선 활동 결과서
	(총족기준)	• 데이터 표준관리를 적용하여 수집된 품질 측정 자료는 이런 비즈니스 환경을 보완할 수 있는 경영 도구로 활용됨을 보장해야 함		

I

II

III

데이터 품질관리 기준

IV

V

5.3.6 구축단계 프로세스 - 데이터 구조관리 프로세스 수준 평가 항목

단계	구축		평가	증빙자료
관련프로세스	데이터 구조			
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 중복 데이터에 대한 목록 정의		
	평가항목	1. 데이터 구조의 정의 및 산출물을 관리합니까? - 데이터 구조의 주요 데이터 목록(테이블 정의서, 컬럼 정의서)을 포함하여 데이터 구조를 정의하고 산출물 관리를 해야 함		<ul style="list-style-type: none">테이블 정의서컬럼 정의서
	(충족기준)	<ul style="list-style-type: none">산출물은 테이블정의서, 컬럼정의서에 반드시 포함하여야 함		
	관리기준	2-2 데이터에 대한 처리 절차 정의		
	평가항목	1. 데이터 간의 일관성 보장을 위한 처리 절차가 존재합니까? - 데이터의 일관성 보장을 위해 동기화, 업무규칙 등을 통한 중복 컬럼의 Value 일관성을 검증해야 함 - 데이터의 일관성을 보장하기 위한 처리 절차를 정의하여 적용해야 함		<ul style="list-style-type: none">데이터 구조 절차서 (중복데이터간 데이터값 일관성 여부점검)
		2. 데이터의 원천 데이터를 정의하고, 이에 대한 오너, 사용자 관계 정의가 되어 있습니까? - 데이터 원천 데이터를 정의하고 이에 대한 오너, 사용자 관계를 정의해야 함		<ul style="list-style-type: none">업무분장내역서
		3. 이음동의 컬럼 등 데이터에 대한 목록을 관리합니까? - 이음동의 컬럼 등 데이터에 대한 목록을 관리해야 함		<ul style="list-style-type: none">테이블 정의서컬럼 정의서
	(충족기준)	<ul style="list-style-type: none">데이터 일관성 보장을 위해 데이터 목록관리(시스템명, Primary(원천)(주제영역, 엔티티명, 테이블명, 속성명, 컬럼명), Child(중복)(주제영역, 엔티티명, 테이블명, 속성명, 컬럼명))를 해야 함데이터에 대한 오너, 사용자 관계 등을 정의해야 함		
	관리기준	2-3 데이터 구조를 통하여 일관성을 유지		
	평가항목	1. 표준화된 방법으로 데이터 모델 관리를 하고 있습니까? - 일관성 있는 데이터 구조 표기법을 정의하고 관리해야 함 · 일관성 있는 데이터 구조 표기법이 존재함을 보장해야 함		<ul style="list-style-type: none">데이터 모델링 가이드(지침서)
		2. 데이터 모델이 정의되고 이를 근간으로 물리DB를 생성합니까? - 데이터 모델과 물리 데이터베이스 간에 일치성 보장여부를 확인해야 함 - 물리 DB와 데이터 모델 간에 일치성을 보장해야 함		<ul style="list-style-type: none">물리 데이터 모델 다이어그램 (데이터관계정의서)논리데이터모델 다이어그램
		3. 데이터 모델의 변경 이력이 관리되고, 이에 대한 최신성을 유지합니까? - 최신의 정보를 제공함으로써 이전 이미지 사용에 따른 사용자의 재작업 가능성을 최소화함		<ul style="list-style-type: none">데이터 모델 변경 이력서
	(충족기준)	<ul style="list-style-type: none">데이터 구조와 관련된 산출물(물리데이터 모델 다이어그램, 논리데이터 모델 다이어그램)을 포함하여 데이터 구조를 정의하고 산출물을 관리하고 있어야 함		

단계		구축	평가	증빙자료
관련프로세스		데이터 구조		
수준	성숙도 수준 평가 항목			
3	관리기준	3-1 데이터 구조를 전사적(플랫폼+센터)으로 통합 관리하고 점검 및 개선활동을 수행		
	평가항목	1. 데이터 구조 산출물의 현행화(70% 이상)를 수행합니까? - DB구조 산출물이 실제 DB의 구조에 맞게 현행화하여 관리하고 있음 - 구조 산출물의 현행화(테이블정의서, 컬럼정의서 기준) 비율이 70% 이상 - 측정산식 : · 불일치율(%) = ((미현행화 테이블 수/실제 DB기준 전체 테이블 수)*100 + (미현행화 컬럼수/실제 DB기준 전체 컬럼 수)*100))/2 · 현행화율(%) = 100(%) - 불일치율(%) *비교키 : 테이블(테이블명), 컬럼(영문명+타입+길이) *평가 제외 테이블 유형 : 임시테이블, 미사용 중복테이블, 백업 테이블, 로그기록 테이블 등		• 데이터 구조 현황 분석서 (불일치성 분석내용 등)
	(충족기준)	• 데이터 구조 산출물의 현행화가 70%이상 유지되어야 함		
4	관리기준	4-1 데이터 구조를 전사적으로 통합 관리하고 점검 및 개선활동을 수행(계속)		
	평가항목	1. 데이터 구조 산출물의 현행화(80% 이상)를 수행합니까? - DB구조 산출물이 실제 DB의 구조에 맞게 현행화하여 관리하고 있음 - 구조 산출물의 현행화(테이블정의서, 컬럼정의서 기준) 비율이 80% 이상		• 데이터 구조 현황 분석서
	(충족기준)	• 데이터 구조 산출의 현행화(테이블정의서, 컬럼정의서 기준)비율이 80% 이상 유지됨을 제시해야함		
5	관리기준	5-1 데이터 구조를 전사적으로 통합 관리하고 점검 및 개선활동을 수행(계속)		
	평가항목	1. 데이터 구조 산출물의 현행화(90% 이상)를 유지합니까? - DB구조 산출물이 실제 DB의 구조에 맞게 현행화하여 관리하고 있음 - 구조 산출물의 현행화(테이블정의서, 컬럼정의서 기준) 비율이 90% 이상		• 데이터 구조 현황 분석서
		2. 오류의 근본원인을 파악하여 이를 해결하기 위한 계획을 수립하고 관련 비즈니스 프로세스를 개선합니까? - 오류의 근본원인을 파악하고 해결하기 위한 계획 수립과 관련 프로세스 개선활동을 전개해야 함		• 오류 개선결과서
	(충족기준)	• 데이터 구조 산출물 현행화가 이루어지고 있음 • 데이터 구조 오류의 근본적인 원인을 해결해야 함		

I

II

III

데이터 품질관리 기준

IV

V

5.3.7 구축단계 프로세스 - 데이터 연계관리 프로세스 수준 평가 항목

단계		구축	평가	증빙자료
관련프로세스		데이터 연계관리(데이터 흐름)		
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 데이터 연계(흐름)를 정의하고 관리		
	평가항목	1. 데이터 연계(흐름)을 정의하고 관리합니까? - 데이터 이동에 필요한 소스와 타겟을 정의하고 소스와 타겟 간의 매핑규칙을 정의해야 함 - 데이터 흐름 모니터링에 의한 오류 관리 - 데이터 흐름 오류에 대한 검증 및 조치 절차를 정의해야 함		<ul style="list-style-type: none"> 데이터매핑 정의서 오류검증/조치 절차서
	(충족기준)	<ul style="list-style-type: none"> 소스 대 타겟 매핑 규칙(원천시스템(테이블명, 컬럼명, 데이터 타입, 길이), 목표 시스템(테이블명, 컬럼명, 데이터 타입, 길이), 변환로직)을 정의하고 관리해야 함 자동화 도구(ETL 등)를 활용하여 데이터 흐름 작업을 스케줄링에 의해 수행하고 이행에 대한 모니터링 결과를 바로 확인할 수 있는 도구가 활용되고 있음 		
	관리기준	2-2 데이터 연계(흐름) 주기를 명확하게 정의		
	평가항목	1. 데이터 연계(흐름)에 대해서 체계적으로 매핑을 관리하고, 최신성을 관리합니까? - 데이터 연계(흐름)에 대한 주기와 매핑 정보를 명확하게 정의해야 함		<ul style="list-style-type: none"> 데이터매핑정의서
		2. 데이터 연계(흐름)상 오류발생 시 피드백(Feedback) 또는 재실행 절차를 관리합니까? - 데이터 연계(흐름) 오류사항은 최대한 많은 경우를 미리 예측하고 이에 대한 대비책을 준비(재실행 절차까지 포함) 해야 함		<ul style="list-style-type: none"> 오류검증/조치 절차서
		3. 데이터의 생성 주기를 파악하고, 데이터 흐름의 선후관계를 조율하고 있습니까? - 정의된 각 데이터 마다 데이터의 생성 주기가 파악되어야 함		<ul style="list-style-type: none"> 데이터매핑 정의서 (데이터생성주기, 발생량, 보존기한 관리)
3	(충족기준)	<ul style="list-style-type: none"> 데이터 매핑의 유형을 일대일, 수평통합, 수직통합, 컬럼 분할 등으로 구분하여 세분화하여 관리함 (데이터 연계(흐름) 관련 매핑 정보관리 : (데이터 매핑 유형, 데이터 엔티티명, to-be테이블, To-be데이터 타입, 데이터 길이, AS-IS 테이블명, AS-IS 컬럼명, AS-IS데이터 타입, 데이터 길이, 변환유형, 변환 SQL) 데이터 연계(흐름)상의 오류 발생에 대해서 프로그램을 재실행하는 것뿐만 아니라 해당 오류 데이터 유형에 따라 사후 조치 사항에 대해서 방안을 수립하고 오류 데이터를 정제해야 함 		
	관리기준	3-1 전사적으로 데이터 연계(흐름) 관리 및 활용		
	평가항목	1. 데이터 연계 품질점검 및 개선(1회/분기 이상) 활동을 수행합니까? - 정해진 요건에 따라 데이터 연계가 정확하게 이루어지고 있는지 품질점검을 실시하고 개선사항 발굴 및 조치하고 있음		<ul style="list-style-type: none"> 연계데이터 품질점검 결과서

단계	구축	평가	증빙자료
관련프로세스	데이터 연계관리(데이터 흐름)		
수준	성숙도 수준 평가 항목		
	2. 데이터 연계(흐름) 내역 관리를 합니까? - 데이터 연계(흐름) 이력이나 로그 등을 기록하고 관리하고 있음		• 연계데이터 품질점검 결과서 (시스템 로그 기록)
	3. 전사적으로 데이터 연계(흐름)에 대한 오류 추적을 통제합니까? - 데이터 매핑정보를 활용하여 변경 영향도가 큰 사안에 대한 변경 시 관련 부서 등과 합의하여 변경 통제, 관리해야 함		• 연계데이터 변경이력서
	(충족기준)		<ul style="list-style-type: none"> • 데이터 연계(흐름) 품질점검은 데이터 연계(흐름) 관리 정보의 모든 데이터를 대상으로 실시하는 것이 원칙임 • 품질점검 실시에 따른 값 점검 결과는 증빙자료를 제출하지 않아도 인정
	관리기준	3-2 전사적인 데이터 연계(흐름) 주기 정의	
	평가항목	1. 전사적인 차원에서 데이터 연계(흐름)에 대해 체계적으로 매핑을 관리하고 최신성을 관리합니까? - 데이터 연계(흐름)관리는 이기종 사이의 데이터 흐름뿐만 아니라 하나의 시스템 안에서의 데이터 연계(흐름)를 관리해야 함(전사 차원에서 모든 데이터의 연계(흐름) 및 매핑 정보를 관리할 수 있음)	• 데이터매핑정의서
		2. 데이터 연계(흐름)의 오류가 발생할 경우, 이에 대한 재실행 및 처리를 위해 전사적 기준을 마련하여 관리합니까? - 각 업무 분야별 데이터 연계(흐름)상 오류가 발생할 경우에 상세한 기준과 절차를 수립하고 시행해야 함 - 업무 영역별로 일관된 기준과 절차를 수립하여 관리하기 위해서는 전사차원에서의 데이터 연계(흐름)상의 오류에 대한 기준과 표준 절차를 수립하고 관리해야 함	• 오류검증/조치 절차서
		3. 전사적으로 데이터 연계(흐름) 주기를 파악하고, 이에 대한 선후, 부하 등을 조정하고 관리합니까? - 전사차원의 데이터 연계(흐름)를 주기적으로 파악하고, 선후 및 부하 등을 조정하고 관리해야함	• 데이터매핑정의서
		4. 전사적인 데이터 연계(흐름)의 선후 관계가 조율되고, 우선 순위에 따라 자원을 배분하고 있습니까? - 사용자 UI프로그램으로부터 데이터베이스로 데이터가 쓰여 지고, 데이터베이스에 있는 데이터가 사용자 UI 프로그램을 통하여 화면에 보여지는 것을 포함	• 데이터매핑정의서 (데이터 연계 선후 관계 정의)
	(충족기준)	<ul style="list-style-type: none"> • 전사 차원에서 데이터 구조가 선행해서 관리되어야 함 • 각 데이터의 발생주기, 분산 및 중복 데이터에 대한 매핑 정보들이 관리되어야 함 • 소스 데이터와 타깃 데이터 사이의 매핑정보가 전사 차원에서 관리되고 각 매핑정보에 의해 생성된 데이터 연계(흐름) 관련 프로그램이 관리되어야 함 	

단계		구축	평가	증빙자료
관련프로세스		데이터 연계관리(데이터 흐름)		
수준	성숙도 수준 평가 항목			
4	관리기준	4-1 데이터 연계(흐름) 관련 품질지표를 관리하고 이를 향상하기 위한 활동을 전사적으로 수행		
	평가항목	1. 조치 가능한 데이터 연계(흐름) 품질점검 및 개선(1회/분기 이상) 활동을 수행합니까? - 조치 가능한 데이터 흐름의 품질 점검 및 개선 활동을 수행하고 있음 - 개선사항 발굴 결과, 개선 내역 등을 의사결정권자에게 보고하고, 관련 이해관계자와 공유하고 있음		<ul style="list-style-type: none"> 연계데이터 품질점검결과서
		2. 데이터 연계(흐름) 담당자 및 이해관계자가 참여하여 데이터 연계(흐름) 품질 향상을 위한 정기적인 의사소통 실시(1회/분기 이상) 합니까? - 데이터 연계(흐름)의 효율적 관리를 위해 연계(흐름) 주기, 방법, 연계 기술 등 데이터 연계(흐름) 관리 체계를 지속적으로 개선하고 있음		<ul style="list-style-type: none"> 연계데이터 품질점검 결과서 (의사소통활동실적, 연계(흐름)주기, 방법, 기술등에 대한 개선내역)
		3. 데이터 연계(흐름) 측정 결과 발생한 오류를 유형화하여 분류하고, 오류에 대한 설명, 발생요인, 해결 내역 등을 관리하며 이에 대한 지식을 축적하여 활용합니까? - 데이터 연계(흐름) 측정 결과에 대한 해결내역 관리와 이에 대한 지식을 축적하여 활용해야 함		<ul style="list-style-type: none"> 연계데이터 품질점검결과서
	(총족기준)	<ul style="list-style-type: none"> 데이터 연계(흐름) 품질개선 활동에 대한 공식화된 증빙자료를 제출해야 함 데이터 연계(흐름)에 대한 오류가 없어 개선활동을 수행할 수 없는 경우, 이 사실을 확인할 수 있는 공식화된 증빙을 제출해야 함(보고서 등) 데이터 연계(흐름) 측정결과 발생한 오류에 대한 해결내역을 관리하고, 이에 대한 지식을 축적하여 활용 및 공유함 		
5	관리기준	5-1 데이터 연계(흐름) 관련 품질지표를 관리하고 이를 향상하기 위한 활동을 전사적으로 수행 (계속)		
	평가항목	1. 조치 가능한 데이터 연계(흐름) 품질개선(1회/분기 이상) 활동을 수행합니까? - 조치 가능한 데이터 연계(흐름)의 품질개선 활동을 수행하고 있음 - 개선사항 발굴 결과, 개선 내역 등을 의사결정권자에게 보고하고, 관련 이해관계자와 공유하고 있음		<ul style="list-style-type: none"> 연계데이터 품질점검결과서
	(총족기준)	<ul style="list-style-type: none"> 조치 가능한 데이터 연계(흐름) 품질개선 활동이 1회/분기 이상 실시함을 증적자료로 제시되어야 함 		

5.3.8 구축단계 프로세스 - 데이터 보호관리 프로세스 수준 평가 항목

단계		구축		평가	증빙자료
관련프로세스		데이터 보호관리			
수준	성숙도 수준 평가 항목				
2	관리기준	2-1 기본적인 수준의 DB 접근통제 및 보호체제 정비			
	평가항목	1. DB서버에 접속할 수 있는 사용자에 대해 정의하고 있습니까? - 데이터베이스 사용자는 관리대상 DBMS에 접속할 수 있는 사람 혹은 서비스를 의미하며 사용자 접속권한이 부여되어야 함			• 업무분장 내역서
		2. DB 사용자 별로 개인을 식별할 수 있는 방법을 정의하고 있습니까? - 개인을 식별하기 위하여 DBMS내에 개인별로 계정을 생성하여 관리해야 함			• 식별 인증 확인서
		3. 로그인 통제 규칙을 정의하고 있고, 통제 활동을 수행합니까? - DBMS에 로그인하기 위해서는 반드시 DBMS내에 생성된 계정 및 패스워드를 통해 통제활동을 수행함			• 사용자별 로그인 권한 조사서
		4. 데이터 보호에 대한 소유자(오너십)가 지정되어 있습니까? - 각 데이터 요소별 데이터 담당자가 명문화 되어 있으며 조직 개편시 이에 대한 현행화 작업을 수행하고 있음			• 업무분장 내역서
		5. 데이터 보호계획을 수립(백업계획 포함) 합니까? - 기관의 데이터베이스 보호계획을 수립하고 있음			• 데이터 보호계획서
	(충족기준)	• 데이터 보호체제가 정의되어 있고 통제활동 수행 및 데이터 보호에 대한 오너십이 지정되어 있음 • 데이터베이스 보호계획(백업계획 포함)이 수립되지 않았거나 부분적으로 수립되어 4등급의 요건을 충족하지 못하는 경우 5등급 부여			
	관리기준	2-2 관리적인 보안 체계 구축			
	평가항목	1. DBMS를 관리하기 위한 보안 절차나 규칙 등을 관리합니까? - 보호절차나 규칙 등을 관리해야 함			• 보호관리 지침서
		2. DB보호 관리자에 대한 역할 및 책임을 정의하고 있습니까? - 보호관리자의 보호관련 점검 내용을 역할분장내역서 등에 정의하고 있음			• 업무분장 내역서
		3. 심화된 DB 접근통제를 하고 있습니까? - 사용자는 해당 업무를 수행하기 위해 필요한 최소한의 권한만을 갖도록 통제해야 함			• 업무분장 내역서
		4. 데이터 소유자에 의해 데이터의 통제(등록, 변경, 삭제)를 수행합니까? - 데이터 소유자가 직접 데이터를 변경하거나, 통제를 통해 데이터가 관리되어야 함			• 업무분장 내역서
		5. 전사차원의 데이터 보호계획을 수립(침해 사고에 대한 복구 계획을 수립 포함) 합니까? - 침해사고 발생 시, 이에 대한 복구계획을 수립하고 있음 - 전사차원의 데이터 보호계획을 수립하고 공식화하고 있음			• 데이터보호 계획서 및 공식화 내역 (침해사고 복구 계획 포함)

I

II

III

데이터 품질관리 기준

IV

V

단계		구축	평가	증빙자료
관련프로세스		데이터 보호관리		
수준	성숙도 수준 평가 항목			
3	(충족기준)	- 전사 차원의 데이터 보호계획은 반드시 별도의 계획으로 작성되어야 하는 것은 아니며, 다른 상위 계획에 포함되어도 인정		
	관리기준	3-1 보호감사를 수행하고 개인정보에 대하여 암호화 등의 추가적인 보호		
	평가항목	1. 데이터 보호 계획 대비 추진실적 점검 및 조치(70% 이상) 활동을 수행합니까? - 데이터 보호계획 대비 추진 실적을 점검하고, 점검에 따른 조치 활동(조치계획, 조치실적 등)을 수행하고 있음 - 데이터 보호계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치활동을 70% 이상을 수행 - 측정산식 : · 계획 대비 실적 달성률(%) = (추진계획 수행 건수 / 데이터 보호계획의 추진과제 계획 건수)*100		• 데이터보호 실적 결과서
		2. DB접근통제를 수행한 로그에 대한 감사를 진행합니까? - 정기적으로 로그에 대한 보고서를 작성한 후 내부적으로 검토하고 데이터 보호 위반 사건이 발생하거나 평상시와 다른 형태의 접근이 있는지 점검(데이터 보호계획서에 점검주기 등이 구체적으로 제시되어야 함)		• 로그에 대한 보호감사 결과서
		3. 비암호화 대상 개인정보에 대해 마스킹 등의 세밀한 DB접근 통제를 수행합니까? - 비암호화 대상 개인정보에 대해 세밀한 DB접근 통제를 실시해야 함 (마스킹이란 데이터 조회 시 권한이 없는 사용자가 중요한 정보를 조회하는 경우 일부 데이터를 '*'값으로 변경하여 원래의 내용을 볼 수 없도록 보호하는 기술임)		• 조회 결과 변조 마스킹
	(충족기준)	<ul style="list-style-type: none"> 데이터 보호 계획서에 포함된 추진과제(백업계획, 침해사고 복구계획, 정보보안감사 계획 등)에 대한 수행여부의 점검결과를 확인해함 계획 대비 실적 달성률은 70% 이상 달성되었음을 증적으로 제시되어야 함 		
4	관리기준	4-1 데이터 보호 이행점검 및 조치(계속)		
	평가항목	1. 데이터 보호계획 대비 추진실적 점검 및 조치(80%이상) - 데이터 보호계획 대비 추진실적을 점검하고, 점검에 따른 조치 활동(조치계획, 조치실적 등)을 수행하였음 - 데이터 보호계획에 있는 주요 추진과제 건수를 기준으로 점검 및 조치활동을 80% 이상 수행		• 데이터 보호 추진실적 점검 결과서
	관리기준	4-2 취약점에 대한 분석 및 조치를 수행		
	평가항목	1. DBMS에 대한 취약점을 분석하여 불필요한 계정을 Lock하고, 권한이 취약한 함수 및 패키지 등의 사용여부를 점검 및 조치합니까? - 네트워크 범위에 존재하는 정보자산을 파악하는 Information Gathering과 데이터 보호를 검증 할 수 있는 모의해킹 및 내부		• 취약점 분석 및 조치결과서

단계	구축		평가	증빙자료
관련프로세스	데이터 보호관리			
수준	성숙도	수준 평가 항목		
		보호감사 과정을 통하여 취약점을 분석 후 검출된 취약점을 제거, 개선안 제시, 보고 작업 등을 조치해야 함		
	(충족기준)	• 데이터 보호계획 대비 실적에 포함하여 점검하고 특히 취약점에 대한 분석 및 조치 수행 결과를 확인하여 수행되고 있음을 보장해야 함		
	관리기준	4-3 인가된 관리자에 의한 정보 유출 방지		
	평가항목	1. 시스템 관리자, DB관리자, 보안 관리자의 역할을 분담하고 이중화된 유출 차단 정책 및 중요 데이터를 선정하며 적절한 수준의 암호화를 수행합니까? - 인가된 관리자(시스템, DB, 정보보안)의 역할 분담실시, 중요 데이터를 선정하며 적절한 수준이 암호화를 수행해야 함		• 인가된 관리자에 의한 정보 유출 방지와 암호화 결과서
	(충족기준)	• 인가된 관리자에 의해 정보유출 방지에 대한 활동 점검 결과를 제시하여야 함 (책임과 권한 정의와 함께 확인 필요)		
5	관리기준	5-1 데이터 보호 이행점검 및 조치(계속)		
	평가항목	1. 데이터 보호 계획 대비 추진실적 점검 및 조치(90% 이상)를 수행합니까? - 데이터 보호계획 대비 추진 실적을 점검하고, 점검에 따른 조치 활동(조치계획, 조치실적 등)을 수행하였음(계속) - 데이터 보호계획에 있는 주요 추진과제 진수를 기준으로 점검 및 조치활동을 90% 이상을 수행		• 중장기 데이터 보호 추진실적 결과서
	(충족기준)	• 당해 연도 중장기 데이터 보호계획을 수립하고 첫해로써 당해 과제를 수행하는 경우도 인정 • 중장기 데이터 보호계획을 수립하고 개선과제의 이행실태를 점검, 개선조치 활동을 90% 이상 수행하고 있음		
	관리기준	5-2 계층적 위협관리 체계 구축 및 보호 수준에 대한 평가		
	평가항목	1. 계층적 위협관리 체계 구축 및 보호수준에 대한 주기적으로 평가를 실시합니까? - DB 사용내역을 주기적으로 분석하여 이상 징후에 대비하여, 직접적인 DB 접근 경로를 차단해 이상 징후 발견 시 자동으로 모든 접속을 차단할 수 있는 체계를 유지해야 함 - 주기적으로 데이터 보호에 관련된 지침 및 절차에 대해 평가를 수행하며, 일련의 보호 행위들이 조직의 데이터 보호 목적에 부합하는지 검증해야 함		• 보호수준평가 결과서
	(충족기준)	• 당해 연도 중장기 데이터 보호계획을 수립하고 첫해로써 당해 과제를 수행하는 경우도 인정 • 중장기 데이터 보호 계획을 수립하고 개선과제의 이행실태를 점검하고 개선조치 활동을 90% 이상 수행하고 있음 • 계층적 위협관리 체계 구축과 주기적으로 보호수준에 대한 데이터 보호 수준평가를 실시함		

5.3.9 운영단계 프로세스 - 데이터 품질진단 및 개선관리 프로세스 수준 평가 항목

단계	운영		평가	증빙자료
관련프로세스	데이터 품질진단 및 개선관리			
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 데이터 품질진단 수행 요건 정비		
	평가항목	1. 품질진단을 사전에 준비하고 있습니까? - DB 품질진단 수행에 필요한 대상 테이블 및 컬럼을 식별하고 진단 대상 전체 컬럼에 대한 진단 기준을 정의하고 있음 · 품질진단 기준은 정의된 품질진단 대상에 매핑되어서 관리되어야 함. 즉 데이터 품질진단 대상 컬럼에 매핑된 품질진단 기준이 식별되어야 함 * 진단 대상 : 진단기준에 해당하는 전체 테이블 및 컬럼		• 데이터 품질진단 대상 정의서
	(충족기준)	• 진단기준은 정의된 품질진단 대상(컬럼)에 매핑 되어서 관리되어야 함		
3	관리기준	3-1 데이터 품질진단 및 개선		
	평가항목	1. 품질진단 실시(70% 이상) 활동을 수행합니까? - 정의된 품질진단 대상 및 기준을 활용한 품질진단 활동을 실시하고 있음 · 품질진단 실시는 정의된 진단대상 컬럼의 70% 이상을 실시하여야 함 - 측정산식 : · 진단 이행률(%) = (당해 연도 진단수행 컬럼 건수 / 당해 연도 진단 대상 DB 컬럼 건수) * 100		• 데이터 품질진단 실시 내역
		2. 품질진단 개선 계획을 수립하고 있습니까? - 데이터 품질진단에 따른 오류의 원인을 분석하고, 오류 정제를 위한 개선계획을 수립하고 있음 - 개선계획은 당해 연도 조치 가능한 개선사항(단기)과 차년도 이후 조치 가능한 개선사항(중장기)을 구분하여 정의 - 당해 연도 계획에는 실제 오류 데이터 정비를 통한 오류 개선율 (%) 제고 활동이 포함되어야 함		• 데이터 품질 개선 계획(단기, 중장기 개선 구분)
		3. 개선 과제 이행률이 70% 이상 달성되고 있습니까? - 수립된 개선계획에 따라 조치 가능한 개선 활동을 수행하고 있음 · 당해 연도 개선계획 대비 이행률은 70% 이상이 되어야 함 - 측정산식 : · 개선 이행률(%) = (당해 연도 개선수행 과제 건수 / 당해 연도 개선계획 과제 건수)*100		• 데이터 품질 개선 계획(단기, 중장기 개선 구분)
	(충족기준)	• 데이터 품질 개선계획은 공식화된 증거를 제출해야 함 • 개선계획에는 오류 데이터에 대한 실제 오류 개선을 제고 노력(활동)이 반영되어야 함 • 해당 기관에서 보유한 도구를 사용해도 무방함		
4	관리기준	4-1 데이터 품질 진단 및 개선(계속)		
	평가항목	1. 데이터 품질진단을 80% 이상 실시합니까? - 정의된 품질진단 대상 및 기준을 활용한 품질진단 활동을 실시하고 있음 · 품질진단의 실시는 정의된 진단 대상 컬럼의 80% 이상 실시해야 함		• 데이터 품질진단 실시 내역

단계		운영	평가	증빙자료
관련프로세스		데이터 품질진단 및 개선관리		
수준	성숙도 수준 평가 항목			
		2. 개선과제 이행률을 80% 이상 달성하고 있습니까? - 수립된 개선계획에 따라 조치 가능한 개선활동을 수행하고 있음 · 당해 연도 개선 계획 대비 이행률은 80% 이상이 되어야 함		<ul style="list-style-type: none"> 데이터 품질진단 개선 조치 결과서
5	관리기준	5-1 데이터 품질 진단 및 개선(계속)		
	평가항목	1. 품질진단을 90% 이상 실시하고 있습니까? - 정의된 품질진단 대상 및 기준을 활용한 품질진단 활동을 실시하고 있음 · 품질진단 실시는 정의된 진단대상 컬럼의 90% 이상을 실시하여야 함		<ul style="list-style-type: none"> 데이터 품질진단 개선 조치 결과서
		2. 개선과제 이행률을 90% 이상 달성하고 있습니까? - 수립된 개선계획에 따라 조치 가능한 개선활동을 수행하고 있음 · 당해 연도 개선 계획 대비 이행률은 90% 이상이 되어야 함		<ul style="list-style-type: none"> 데이터 품질진단 개선 조치 결과서

5.3.10 운영단계 프로세스 - 오류 데이터 관리 프로세스 수준 평가 항목

단계		운영	평가	증빙자료
관련프로세스		오류 데이터 관리		
수준	성숙도 수준 평가 항목			
	관리기준	2-1 데이터 품질 오류율		
2	평가항목	1. 품질진단 기준에 따른 품질 정합률이 98.5% 미만 - 품질진단 기준에 따라 품질진단을 실시하고, 그 결과 데이터 품질 정합률이 98.5% 미만(Low Class)		<ul style="list-style-type: none"> 데이터 품질진단 테이블 선정 내역 데이터품질진단항목 기준선정내역 데이터품질진단결과내역
	관리기준	3-1 데이터 품질 오류율(계속)		
3	평가항목	1. 품질진단 기준에 따른 품질 정합률이 98.5% 이상 - 품질진단 기준에 따라 품질진단을 실시하고, 그 결과 데이터 품질 정합률이 98.5% 이상(Middle Class)		<ul style="list-style-type: none"> 데이터 품질진단 테이블 선정 내역 데이터품질진단항목 기준선정내역 데이터품질진단결과내역
	관리기준	4-1 데이터 품질 오류율(계속)		
4	평가항목	1. 품질진단 기준에 따른 품질 정합률이 99.0% 이상 - 품질진단 기준에 따라 품질진단을 실시하고, 그 결과 데이터 품질 정합률이 99.0% 이상(High Class)		<ul style="list-style-type: none"> 데이터 품질진단 테이블 선정 내역 데이터품질진단항목 기준선정내역 데이터품질진단결과내역
	관리기준	5-1 데이터 품질 오류율(계속)		
5	평가항목	1. 품질진단 기준에 따른 품질 정합률이 99.97% 이상 - 품질진단 기준에 따라 품질진단을 실시하고, 그 결과 데이터 품질 정합률이 99.97% 이상(Ace Class)		<ul style="list-style-type: none"> 데이터 품질진단 테이블 선정 내역 데이터품질진단항목 기준선정내역 데이터품질진단결과내역

I

II

III

데이터 품질관리 기준

IV

V

5.3.11 활용단계 프로세스 - 데이터 요구사항관리 프로세스 수준 평가 항목

단계	활용		평가	증빙자료
관련프로세스	요구사항관리			
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 사용자 요구사항 수집		
	평가항목	1. 업무 수행을 위해 사용자의 데이터 요구사항을 충분히 수집하고 있습니까? - 요구사항 수집이 단순히 구두로 이루어지고 있는지, 체계적인 시스템 기능 개선 요청과 수집 절차에 따라 이루어지고 있는지 수집된 요구사항 내역을 관리하고 있는지 점검해야 함 - 내외부 사용자를 대상으로 품질오류 신고, 요구사항 등을 신청 접수할 수 있는 창구를 운영하고 있음		<ul style="list-style-type: none">요구사항수집 내역서신고 접수창구 운영
		2. 사용자의 데이터 요구사항에 대한 처리 현황을 관리합니까? - 수집된 요구사항에 대해서 처리 현황이 관리되고 있음 - 각 요건이 현재 어느 상태에 있는지 관리(우선순위에 따라 처리) - 협의과정 또는 구현과정에서 발생하는 중요한 문제점 및 이슈에 대해서도 관리 · 사용자 요구사항 수집 예시(시스템명, 작성일자, 업무영역, 업무 기능, 요구사항 명, 설명, 상태, 중요도, 난이도, 안정성, 관련 부서, 발의자, 근거문서, 적용방법(반영내역) 등)		<ul style="list-style-type: none">요구사항 현황관리내역서
		3. 사용자의 다양한 데이터 요구사항을 유형별로 분류하여 대처합니까? - 사용자의 다양한 데이터 요구사항을 유형별로 분류하고 대처하고 있음		<ul style="list-style-type: none">요구사항 현황관리내역서
		4. 사용자 요구사항에 대한 처리 현황을 기록하고, 대응 시 문제점을 파악하고 있습니까? - 요구사항이 어떻게 이루어지고 있는지 점검해야 함 - 형식적이거나 급조한 문서 생성여부, 실제 적극적으로 요구사항 관리를 수행하는지 점검해야 함		<ul style="list-style-type: none">요구사항처리 현황, 문제점 파악 현황
	(총족기준)	<ul style="list-style-type: none">신고 접수 창구를 운영하고 있어도 신고 접수 및 처리 기준을 정의하지 않아 수준2의 요건을 충족하지 못한 경우 수준 1 부여		
	관리기준	2-2 사용자 요구사항에 대한 처리기준 정의 및 대응절차를 정형화		
평가항목	1. 데이터와 관련된 사용자 요구사항 뿐만 아니라 어플리케이션, 비즈니스 등에 사용자 요구사항을 정의하여 관리합니까? - 요건에 관계된 다양한 관계자들이 충분히 요건을 이해할 수 있도록 분류하여 관리 - 내외부 사용자 대상으로 품질오류 신고 및 요구사항을 신청 접수할 수 있는 창구를 운영하고 있음 - 수요자 요구사항 처리 절차 및 기준을 정의하고 있음		<ul style="list-style-type: none">요구사항 정의서	
	2. 표준화된 구현 계획에 의해 사용자 요구사항을 대응하고 있습니까? - 사용자의 개선·수정요구사항은 작업 표준화가 수립되어 수행 일정 등이 관리되고 있음		<ul style="list-style-type: none">요구사항대응 내역서	

단계	활용		평가	증빙자료
관련프로세스	요구사항관리			
수준	성숙도 수준 평가 항목			
3		- 작업에 필요한 점검사항 및 협조부서와의 조율 내역 및 계획 등이 관리되고 있음		
		3. 요구사항에 대한 변경처리를 포함하여 처리 결과를 피드백 하는 절차를 정의하고 있습니까? - 사용자 요건을 접수하고, 요건에 만족하는 기능을 사용자가 확인 하기까지 전반적인 과정이 절차화·문서화 되었는지 확인함		• 요구사항 추적 매트릭스
	(총족기준)	• 당해 년도 접수창구를 만든 후 접수 내용이 한 건도 없어서 후속 활동을 할 수 없는 경우 2수준 부여 • 신고 접수 및 처리 절차는 공식화되지 않은 증적 자료도 인정 • 데이터 포털을 활용할 경우 접수창구 인정 • 요구사항의 신고 접수 및 변경처리에 대한 제반 활동을 수행하고 피드백(후속처리)하는 증적을 제출 또는 확인할 수 있어야 함		
	관리기준	3-1 전사적으로 사용자 요구사항을 파악하고 조율		
3	평가항목	1. 사용자 요구사항에 대한 변경 영향도를 분석하는 절차가 존재 하고, 변경 영향 분석을 전사적으로 수행합니까? - 변경된 요청내역을 토대로 변경에 따른 영향이 미치는 설계서 및 어플리케이션, 데이터베이스 등을 도출하고, 그 결과로 변경 영향 도 분석서를 작성하는 절차와 방안을 마련해야 함		• 변경 영향도 분석서
		2. 사용자 요구사항에 대해 전사적인 대응 절차를 정의하고 관리합니까? - 요구사항 접수시점부터 검토, 조율, 개발, 테스트 및 검수단계까지 요건을 제출한 사용자가 진행상황을 알 수 있도록 조치해야 함 - 수요자 요구사항 처리 절차 및 기준을 준수하고 있으며 처리결과를 신청자에게 통지하고 있음		• 요구사항 대응 절차서
		3. 전사적으로 사용자 요구사항을 조율하는 공식화된 기능이 존재 하여 사용자 요구사항을 통합하여 일관성 있게 처리합니까? - 개별적으로 수집된 요건을 전사차원이나 장기적인 측면에서 검토 하고 조율하는 작업이 수행되어야 함 - 전사적인 측면에서 관리하는 조직(기능)에는 조직의 업무를 전반 적으로 잘 이해하고 개선 요구사항이 왜 발생했는지 충분히 이해 하고 있는 담당자의 참여 보장		• 업무분장 내역서
		4. 사용자 요구사항에 대한 우선순위를 관리하며 사용자 요구사항에 대한 신규, 변경 등에 대한 영향도를 평가합니까? - 접수된 요구사항의 업무적인 시급성, IT 및 데이터 요소들간의 영향도를 검토하여 우선순위를 재조정하고 있음		• 요구사항 우선 순위 정의서
(총족기준)	• 사용자 요구사항이 기존 운영되는 시스템에 미치는 영향도를 분석하는 절차를 갖추어야 함			
4	관리기준	4-1 사용자 요구사항 분석 및 개선활동 전개		

I

II

III

데이터 품질관리 기준

IV

V

단계	활용		평가	증빙자료
관련프로세스	요구사항관리			
수준	성숙도 수준 평가 항목			
	평가항목	1. 데이터 품질오류 분석 및 문제점을 개선하고 있습니까? - 수요자 신고 접수 내역 및 처리 결과를 분석하여 수요자 요구사항 파악 및 문제점을 발굴하고 개선하는 활동을 수행하고 있음(연간 1회 이상)		• 개선조치활동 결과서
		2. 요구사항 관련 품질지표를 이용한 분석결과를 관리합니까? - 요구사항 관련 품질지표를 이용하여 각 항목에 대한 측정 결과를 분석하고 관리해야 함 - 분석결과를 이용하여 품질 개선을 위한 활동의 방향을 정할 수 있고, 또 품질지표를 개선하기 위한 자료로 활용하고 있음		• 개선조치활동 결과서
		3. 데이터 품질 만족도 조사 및 요구사항을 조사 및 분석합니까? - 수요자 데이터 이용실태(사용자 만족도 조사, 활용 실태조사, 활용성이 높은 데이터 수요조사)를 매년 조사 분석하고 있음(최근 2년간)		• 성과관리 결과서
	(총족기준)	• 데이터 품질오류 분석 및 문제점 개선은 공식화된 증거를 제출해야 함 • 데이터 품질만족도 및 요구사항 조사 분석은 공식화된 증거를 제출해야 함		
	관리기준	4-2 요구조건의 내부동향 파악		
	평가항목	1. 요구조건의 변경 내역이 분류되고 내부 요건의 변경 성향을 파악하여 요건 대응에 활용합니까? - 요구조건 변경내역에 대한 변경 성향을 파악하여 요구조건 대응에 활용되어야 함		• 개선조치 활동 결과서
(총족기준)	• 요구조건에 대한 제반 활동내역 및 분석 결과를 기반으로 내부동향을 파악하고 활용하고 있음			
5	관리기준	5-1 요구사항 관리 활동이 직접적인 경영개선 활동으로 연계되어 수행		
	평가항목	1. 사용자 요구사항 관련 활동이 정량적인 수치를 근거로 경영개선 활동을 수행합니까? - 수요자 데이터 이용실태 조사결과 등을 통해 파악된 요구사항을 정책(제도, 업무 등)에 반영하고, 경영개선 활동과 연계하여 수행 함(연간 1회 이상)		• 개선조치 활동 결과서
	(총족기준)	• 평가 연도에 해당하는 자료가 존재하지 않을 경우 평가 직전 연도의 연말에 수행된 자료도 인정 • 정책 반영은 주로 연말에 수행되므로 직전 연도의 수행활동으로 미루어 당해 연말에도 수행이 예상되어 인정		
	관리기준	5-2 요구사항의 외부 동향 파악		
	평가항목	1. 업계의 우수사례(Best Practice)를 반영하여 지속적으로 유지, 발전시키고 데이터 활용 기술 및 동향을 파악하여 적용합니까? - 우수사례를 반영하여 지속적으로 데이터 활용 기술 및 동향 파악에 적용해야 함		• 개선조치 활동 결과서
	(총족기준)	• 요구사항에 대해 외부 동향 파악을 위해 최신 기술동향 파악한 자료 제출 시 인정		

5.3.12 활용단계 프로세스 - 사용자 뷰 관리 프로세스 수준 평가 항목

단계	활용		평가	증빙자료
관련프로세스	사용자 뷰 관리			
수준	성숙도 수준 평가 항목			
2	관리기준	2-1 사용자 뷰 구조를 표준화하여 제공		
	평가항목	1. 사용자 뷰의 구성 요소를 표준화하며, 다양한 사용자 지원 기능을 제공 - 화면 설계 표준 등이 존재하고 이를 기반으로 화면설계, 개발 등의 작업을 수행하고 있음		• 화면 설계 표준서
	(충족기준)	• 다양한 사용자 지원 기능을 제공하여야 함(화면 설계 표준 등이 존재)		
	관리기준	2-2 사용자의 데이터 활용을 위한 매뉴얼 지원		
	평가항목	1. 사용자 뷰를 활용하기 위한 다양한 형태의 지원 도구 등을 제공. (매뉴얼, 튜토리얼 등) 또한 이러한 도구들이 최신성을 유지 - 사용자 뷰를 활용하기 위한 다양한 도구를 제공하고 현행화되어야 함		• 지원도구 (매뉴얼, 튜토리얼)
		2. 시스템 사용을 위한 사용자 교육을 주기적으로 시행 - 사용자가 해당 시스템을 잘 활용하기 위해서는 시스템 사용에 대한 적극적인 교육, 홍보, 도구를 제공하여야 함		• 교육 실시 결과서
	(충족기준)	• 사용자 뷰를 활용하기 위해 다양한 형태의 지원 도구 제공해야 하고 최신성을 보장해야 함		
관리기준	2-3 사용자가 직접 데이터를 활용			
평가항목	1. 사용자가 직접 데이터 활용 - 사용자의 직접 데이터 활용을 위하여 데이터 구조가 공유되고, 활용 도구(Tool) 등이 제공되고 있음		• 지원도구 (레포팅 도구 등)	
(충족기준)	• 사용자가 데이터를 직접 활용할 수 있도록 데이터 구조를 공유해야 함			
3	관리기준	3-1 사용자 뷰를 통합된 환경에서 제공		
	평가항목	1. 사용자 뷰의 편리성을 위해 다양한 매체를 통하여 제공. 단, 다양한 매체는 전사적 차원에서 통제(계획) - 사용자 뷰를 다양한 매체를 통하여 제공하며, 각각의 단위 시스템들은 통합된 사용자 뷰를 제공하고 있음		• 사용자 뷰 매체 (화면설계표준서 등)
	(충족기준)	• 사용자 뷰의 편리성을 위해 다양한 매체(화면 설계 표준, 매뉴얼 등)를 제공함을 확인할 수 있어야 함		
	관리기준	3-2 가치 창출을 위한 환경 제공 - 전사적으로 사용자가 직접 데이터를 활용		
	평가항목	1. 전사에서 공통적으로 사용되는 데이터에 대한 통합 뷰를 제공하고, 공통 데이터의 일관성과 최신성을 유지 - 사용자의 정보 접근성을 극대화시키기 위한 노력의 또 다른 축이 전사에서 공통적으로 사용하는 데이터를 통합 뷰를 통하여 제공하고 이의 최신성과 일관성을 지속적으로 유지해야 함		• 화면설계표준서
(충족기준)	• 제공된 데이터 구조 정보를 활용하여 사용자는 직접 데이터를 활용하고 다양한 형태의 분석을 수행할 수 있는 환경을 제공함			

I

II

III

데이터 품질관리 기준

IV

V

단계		활용	평가	증빙자료
관련프로세스		사용자 부 관리		
수준	성숙도 수준 평가 항목			
4	관리기준	4-1 사용자 부 관련 품질지표를 관리하고 이를 향상하기 위한 활동을 전사적으로 수행		
	평가항목	1. 사용자 부 관련 품질지표를 경영지표와 연계하는 활동 수행 - 사용자 부와 관련된 품질지표는 다양함 · 효과성 : 테스트 성공과 실패여부 · 효율성 : 사용자가 검색을 하고자 할 경우 걸리는 평균시간 · 오류횟수 : 장애 요인		· 개선조치활동 결과서
		2. 사용자 부 관련 품질향상을 위해 성과관리 활동을 수행한다. - 데이터 품질지표를 경영상의 평가지표와 연계하여 관리하는 것이 쉬운 것이 아니지만 균형성과 평가제도(BSC) 기법을 인용하기도 함		· 성과관리결과서
	(충족기준)	<ul style="list-style-type: none"> · 사용자 부 관련 품질지표를 경영지표와 연계하는 활동을 수행하고 증적자료를 제출해야 함 · 사용자 부 관련 품질향상을 위해 BCS(재무, 고객, 내부 프로세스, 학습과 성장) 등을 인용한 성과관리 활동 수행결과를 제시해야 함 		
	관리기준	4-2 사용자 활용 패턴을 파악하고 지원		
	평가항목	1. 축적된 데이터를 바탕으로 사용자 활용 패턴을 파악 - 사용자에게 정보를 제공하는 정보시스템은 사용자의 활용 패턴을 이해하고 이를 근거로 사용자 접근성을 향상하기 위한 지속적인 활동을 수행해야 함		· 개인화 된 부 (화면과 기능)
5	(충족기준)	<ul style="list-style-type: none"> · 사용자의 활용 패턴을 이해하고 사용자 접근성을 향상시키기 위한 지속적인 활동이 수행됨을 보장해야 함 (예시로 연령별 포털 구성요소에 대한 선호도 조사를 통해 연령별 개인화된 화면과 기능을 제공하기 위한 기본정보로 활용) 		
	관리기준	5-1 사용자의 데이터 활용이 직접적인 경영개선 활동으로 연계		
	평가항목	1. 사용자는 전사 데이터에 대한 고급 데이터 분석을 수행하고 이를 통한 조직 가치 창출에 활용 - 데이터 분석을 통하여 비즈니스의 새로운 시사점을 도출하고 이를 이용하여 가치를 창출하는 체계를 유지함		· 경영자를 위한 고급 데이터 분석 활용 내역
	(충족기준)	<ul style="list-style-type: none"> · 단순 집계 형태의 데이터를 벗어나 다양한 형태의 고급 데이터 분석 기능을 제공하고 이를 통해 경영 가치 창출에 활용됨을 보장 		
	관리기준	5-2 사용자 부 최적화		
	평가항목	1. 사용자 부의 최적화 - 사용자 부에 대한 최신 기술과 현황을 접목하여 사용자 부를 지속적으로 최적화해야 함		· 사용자 부 최적화
	(충족기준)	<ul style="list-style-type: none"> · 사용자 부에 대한 최신 기술동향을 파악하여 사용자 부를 지속적으로 최적화하는 활동 결과를 제시해야함 (큰 규모의 데이터, 보다 다양한 형태의 데이터, 보다 빠른 생성 속도를 가진 데이터를 관리 및 통제할 수 있고, 이 데이터 분석을 통해 비즈니스의 새로운 시사점을 찾아내고, 가치를 창출하는 체계가 구축되어야 최적화 되었다고 할 수 있음) 		

IV 데이터 품질진단 방법

1장 개요

2장 품질진단 정의(Define)

3장 품질진단 수행(Measure)

4장 진단결과 분석(Analyze)

5장 개선수행(Improve)

6장 품질통제(Control)



01

빅데이터 품질진단 방법 개요



1.1 빅데이터 품질진단 필요성

국내·외적으로 각 국의 정부는 데이터의 중요성을 인식하여 데이터 축적·유통 활성화를 강조하고 있다. 우리 정부는 데이터 산업 활성화 전략을 수립하고 데이터 산업을 혁신성장 전략투자 분야로 선정하였다. 또한 빅데이터 플랫폼 및 센터 구축을 통해 데이터 축적·유통을 활성화하기 위한 데이터 수집, 분석, 유통 체계를 마련 중이다.

그러나 폐쇄적 데이터 축적과 접근성 제약으로 인해 무엇보다 데이터에 대한 인식 개선이 중요하고 빅데이터 품질관리 기반체계 마련이 시급한 실정이다. 이에 고품질 데이터 확보 및 유지를 위한 빅데이터 품질진단 방법론 마련으로 빅데이터 품질관리에 대한 인식 제고 및 품질관리 수준 향상을 도모하고자 한다.

1.2 빅데이터 품질진단 정의

▶ 빅데이터 품질

- “빅데이터의 완전성, 유효성, 일관성, 정확성 등 품질 기준별로 데이터 또는 서비스 수요자에게 유용한 가치를 줄 수 있는 수준”으로 정의할 수 있다.

▶ 빅데이터 품질진단

- 빅데이터 이용 목적에 맞게 정의된 품질지표별 진단기준에 따라 데이터를 프로파일링, 업무규칙, 체크리스트 등의 방법으로 진단규칙을 작성하여 오류 데이터를 탐지한다.
- 탐지된 오류 데이터의 오류 원인 및 유형을 분석하고 오류 데이터를 개선함으로써 데이터 품질 수준을 지속적으로 향상시키는 활동이다.

I

II

III

IV

데이터 품질진단 방법론

V

1.3 빅데이터 품질진단 특징

1.3.1 빅데이터 특성

▶ 빅데이터 기술 특성

- 빅데이터는 데이터 저장구조 및 트랜잭션의 엄격한 제약에서 벗어나 용도에 맞게 다양하고 유연한 자료구조, 빠른 검색과 페타바이트 이상 대량 데이터 처리가 가능하도록 지원하고 있다.
- 빅데이터는 대량 데이터 분석에 의해 통찰력의 발견을 목적으로 하고 있어 빅데이터 시스템의 데이터 품질이 이용 목적에 따라 분석 결과에 미치는 영향도는 업무처리 시스템 데이터 품질보다 상대적으로 적은 편이다.

▶ 빅데이터 분류 특성

- 빅데이터를 분류하는데 데이터를 생산하는 주체에 따라 사람이 생성한 데이터, 프로세스가 생성한 데이터, 기계가 생성한 데이터로 분류하고 있다.
- 또한 빅데이터의 분류를 데이터 유형, 저장 방식에 따라 다양하게 데이터를 분류하여 정의하고 있다.

▶ 빅데이터 품질 특성

- 빅데이터의 역할은 미래의 불확실성, 리스크 등에 대응하기 위한 통찰력, 대응력 등을 얻기 위한 것으로 다양한 유형 및 대량 데이터를 기반으로 사회적, 경제적 통찰력을 찾는 것을 빅데이터 특성으로 볼 수 있다.
- 빅데이터는 분석을 위하여 대량의 데이터 처리 기술을 이용하며 데이터의 정확성보다 속도에 초점을 중점을 두기 때문에 일부 소량 데이터 유·손실이 발생해도 처리 결과에 크게 영향을 주지 않는다.

1.3.2 빅데이터 분류체계

데이터 분류 내용을 기반으로 ‘분석, 문서, GIS’유형을 추가하여 메타데이터, 기준데이터, 로그데이터, 거래데이터, 분석데이터, GIS데이터, 문서데이터, 미디어데이터, 센서데이터, 활동데이터 등 10가지 데이터 유형 분류를 정의한다.

데이터 자료구조 분류 체계는 전통적인 DB 구조인 “RDB”, 빅데이터 처리를 위한 “NoSQL”, 이미지, 영상, 문서 등 비정형 자료인 “Non Structure”로 정의한다.

1.3.3 진단 대상 선정 및 표본진단

대량 빅데이터의 정보 가치 중요도에 따라 진단 대상을 선정하고 대량 데이터에 대한 효율적인 진단 방법을 제시한다.

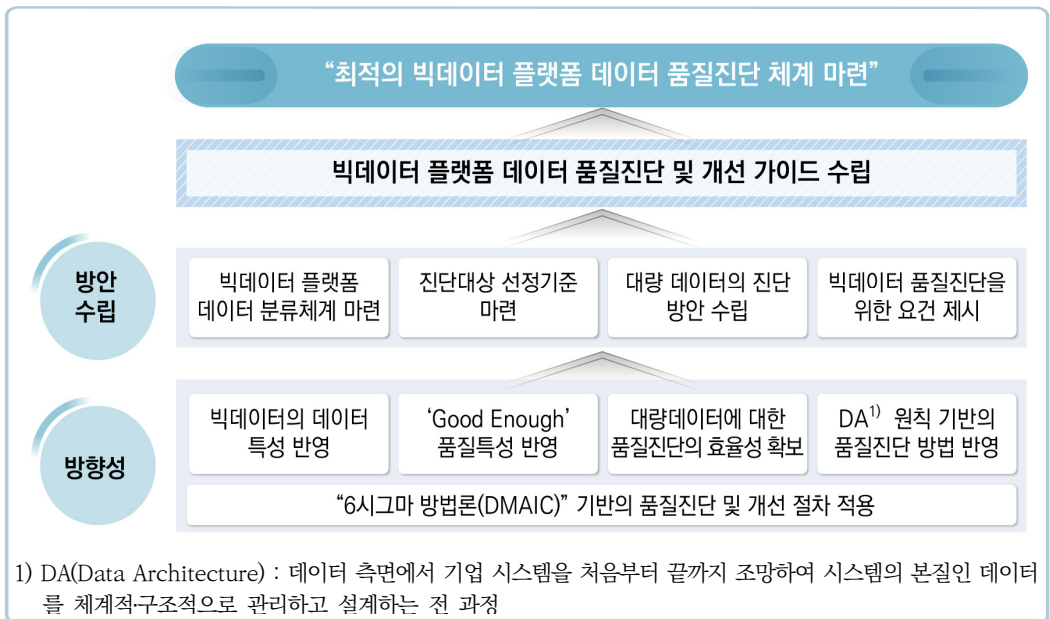
일상적인 빅데이터 품질진단을 효율적으로 수행하기 위해 대량 데이터의 경우 통계적인 기법에 의해 표본을 추출하여 그 대상으로 진단을 수행할 수 있다.

1.3.4 6-시그마 기반 데이터 품질진단 방법론

본 가이드에서 빅데이터의 체계적인 품질진단 및 개선을 위해 “6-시그마 방법론(DMAIC)”기반의 데이터 품질 방법론을 제시한다.

- DMAIC는 Define, Measure, Analyze, Improve, Control 단계의 줄임말로 과거의 경험, 업무에 대한 지식, 통계기법에 의한 근거를 통해, 체계적인 문제해결 과정이다.
- 기존 프로세스에서 발생하는 문제를 해결 혹은 프로세스 성능을 향상시키기 위한 개선 방법론이다.

그림 1 빅데이터 품질진단 방법론 수립 방향성



1.4 빅데이터 품질진단 기법

▶ 프로파일링

정의된 메타 표준정보로 진단규칙 도출이 불가할 경우 “프로파일 분석”이 필요하며 진단대상 항목의 유효값, 유효 범위를 메타정보 분석 또는 통계적 분석으로 오류로 예상되는 항목에 대해 진단 규칙을 도출하여 진단한다.

▶ 업무규칙

업무 관점의 품질이슈나 업무 담당자 인터뷰를 통해 값에 대한 진단 요건을 확보하여 진단 규칙을 도출한다.

▶ 공간정보

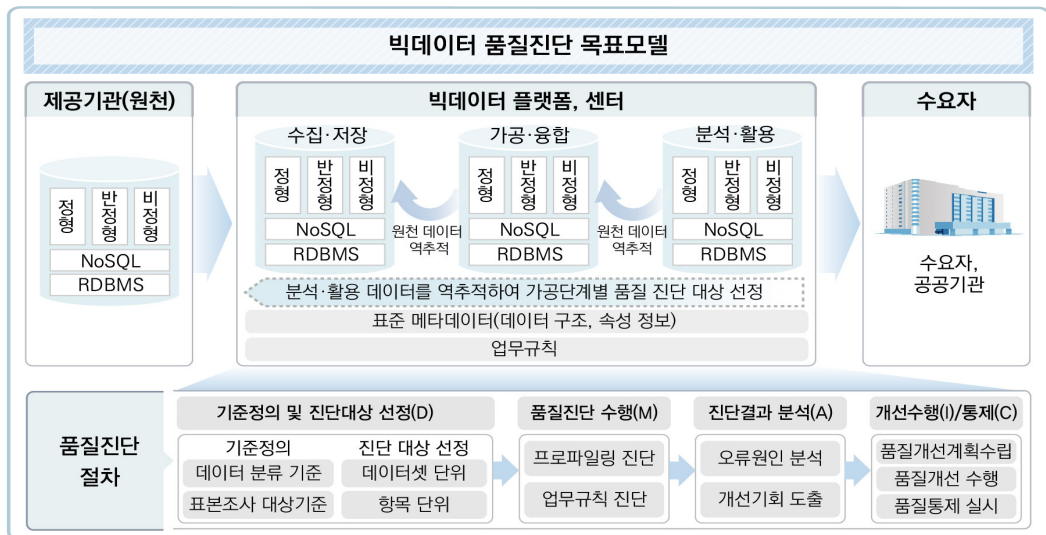
GIS 공간정보의 특성을 고려하여 구조화 기반 공간 진단, 레이어(layer) 기반 공간 진단, 육안검사 등을 통해 공간정보에 대해 잠재적 오류 징후를 발견하기 위한 진단규칙을 도출한다.

1.5 빅데이터 품질진단 목표모델

빅데이터 플랫폼 및 센터 데이터 수집·저장, 가공·융합, 분석·활용 단계 각 영역별로 빅데이터 특성 및 데이터 흐름을 고려한 품질진단 목표 모델을 제시한다.

- 플랫폼 및 센터는 원천으로부터 수집한 데이터를 각 처리 단계 및 영역별로 품질진단을 수행하며 표준 메타데이터 또는 업무규칙에 의해 제공기관(원천) 또는 전 단계의 데이터 일관성 등을 확보한다.

그림 2 빅데이터 품질진단 목표 모델



02

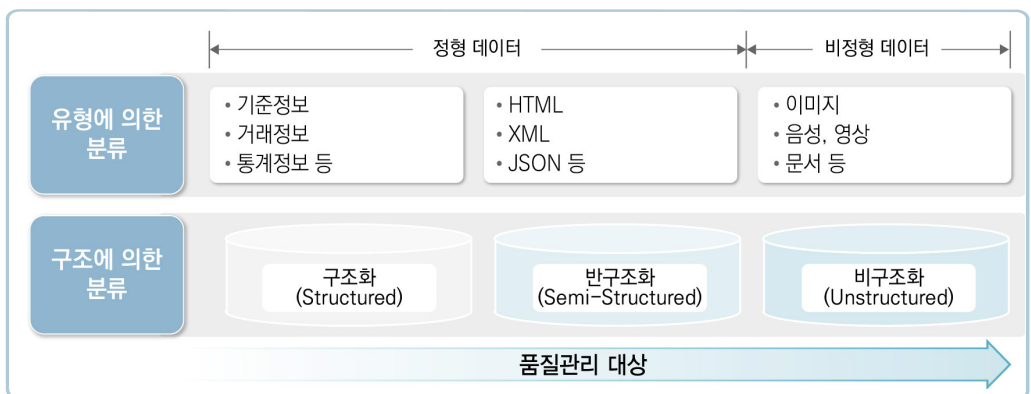
빅데이터 품질특성 및
분류체계 정의

일반적으로 데이터(Data)란, 특정 목적의 활동이나 이벤트로 인해 “발생한 사실(Fact)의 바탕이 되는 자료”를 의미하며, 광의의 개념으로 정보(Information) 또는 지식(Knowledge)과 혼용하여 사용하기도 한다.

데이터는 종이 문서와 같이 디지털화되지 않은 자료도 포함하지만, 데이터의 품질관리 측면에서 볼 때 디지털화되지 않은 데이터는 진단하는 데 있어 진단 도구의 한계로 인하여 품질진단을 수행하는 데 어려움이 따른다.

디지털화된 데이터를 각 데이터가 가지고 있는 특성에 의하여 데이터를 분류하면, 데이터 유형에 따라 기준정보, 거래정보, 통계정보 등으로 분류할 수 있으며, 이와 같은 데이터는 구조화(Structured Data) 형태로 있으며, HTML, XML 등과 같은 유형 데이터는 반 구조화 데이터(Semi-Structured Data) 형태를 보이며, 동영상, 이미지, 소리 등의 유형 데이터는 비구조화 데이터(Unstructured Data) 형태를 갖추고 있다.

그림 3 데이터의 분류 및 품질관리 대상



▶ 공간정보의 구성 및 품질진단 대상

- 공간정보는 지리 정보 중에서 지표 공간 속에 분포하는 현상의 위치나 형태를 나타내는 정보를 공간정보라 한다. 공간정보는 점, 선, 면으로 표현된다. 상점이나 공장의 위치는 점으로, 도로와

하천, 교통·통신망의 이동은 선으로, 산업 단지나 주택 단지와 같이 활동의 영향권은 면으로 표현된다. 공간정보는 위도, 경도, 행정구역의 경계 등으로 표현할 수 있다.

- 공간정보는 지리적 위치를 갖고 있는 대상에 대한 위치자료(spatial data)와 속성자료(attribute data)로 구성되어 있다.

그림 4 공간정보 데이터 구성 및 진단 대상

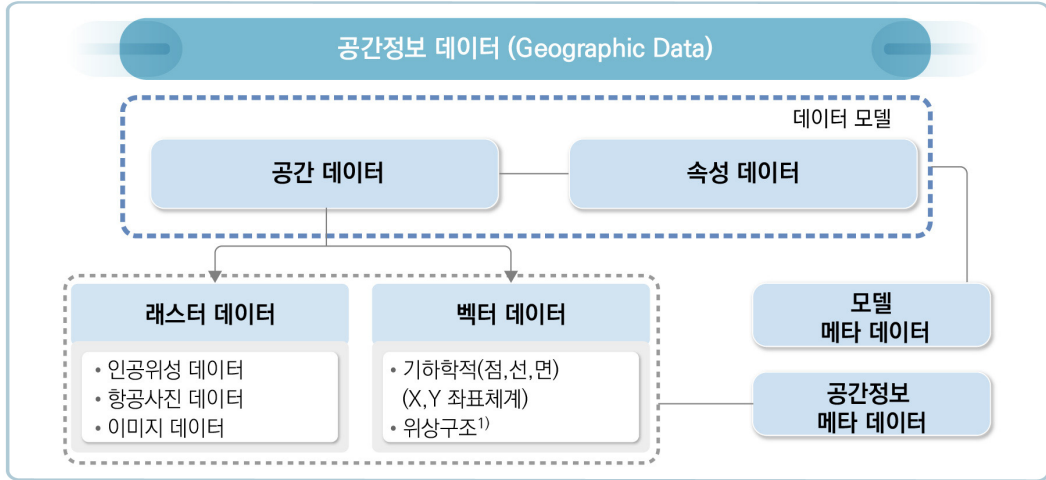


표 1 공간정보 구성요소의 설명

대상 구분	설명	예
공간 데이터 (Spatial data)	<ul style="list-style-type: none"> 지형지물의 공간특성을 표현한 데이터 	<ul style="list-style-type: none"> 벡터데이터, 래스터데이터
래스터 데이터 (Raster data)	<ul style="list-style-type: none"> 셀마다 그 지리적 범위에 관한 특성값을 일정한 수치로 표시하는 격자 형식의 자료 구조를 가지는 자료 	<ul style="list-style-type: none"> 인공위성 영상, CCD 사진기 영상, 또는 항공사진, 정사영상 등
벡터 데이터 (Vector data)	<ul style="list-style-type: none"> 모든 공간정보를 2차원 또는 3차원 좌표값으로 환산하여 점, 선, 면의 형태로 저장한 자료 	<ul style="list-style-type: none"> Shape file 등
속성 데이터	<ul style="list-style-type: none"> 어떤 대상의 주제나 위치 특성을 나타내는 정보 	<ul style="list-style-type: none"> 건물의 층수, 주소, 면적 등
공간정보 메타 데이터	<ul style="list-style-type: none"> 공간정보를 설명하는 정보를 목록화 하여 제공함으로써 사용자가 자료의 획득 및 사용에 도움을 주기 위하여 자료 	<ul style="list-style-type: none"> 좌표계, 해상도, 축척, 도엽기준, 파일형식 등
모델 메타 데이터	<ul style="list-style-type: none"> 속성 데이터의 명명규칙, 표준사전, 표준도메인 등 표준화 관련 정보 	<ul style="list-style-type: none"> 명명규칙, 표준사전, 표준도메인 등

▶ 비정형 데이터의 구성 및 품질진단 대상

비정형 데이터는 문서, 이미지, 음성, 영상 등과 같이 비구조 형태를 가지고 있는 데이터로 비구조 형태의 객체 데이터와 객체 데이터에 대한 객체 메타 데이터로 구성되며, 비즈니스 목적에 따라 관리하는 관리정보를 갖고 있을 수 있다.

그림 5 비정형 데이터 구성 및 진단 대상

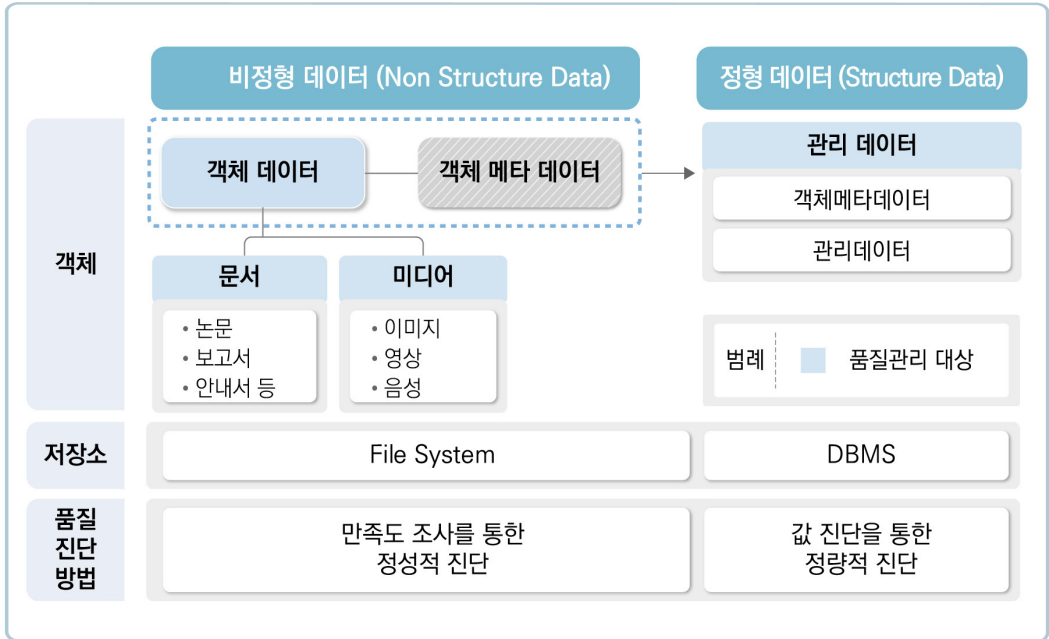


표 2 비정형 데이터 구성요소의 설명

대상 구분	설명	예
객체 데이터	<ul style="list-style-type: none"> 비정형 구조 형태의 데이터로 문서파일, 미디어파일 자체 	<ul style="list-style-type: none"> 문서, 이미지, 음성, 영상 등
객체메타데이터	<ul style="list-style-type: none"> 객체정보를 생성하면서 자동으로 생성되는 정보 	<ul style="list-style-type: none"> 파일명, 파일형식, 해상도, 생성일자 등
관리 데이터	<ul style="list-style-type: none"> 활용을 위해 객체정보를 설명하는 관리데이터항목 및 객체메타데이터 항목 등의 데이터로 구성된 정보 	<ul style="list-style-type: none"> 저장위치, 주제영역, 파일명, 파일형식 등
문서	<ul style="list-style-type: none"> 문서작성기로 생성한 문서 정보(고유의 저장형식을 가짐) 	<ul style="list-style-type: none"> txt, hwp, doc, pdf 등
미디어	<ul style="list-style-type: none"> 이미지, 음성, 영상 등의 미디어 정보(고유의 저장형식을 가짐) 	<ul style="list-style-type: none"> jpg, gif, png, mp3, mp4 등

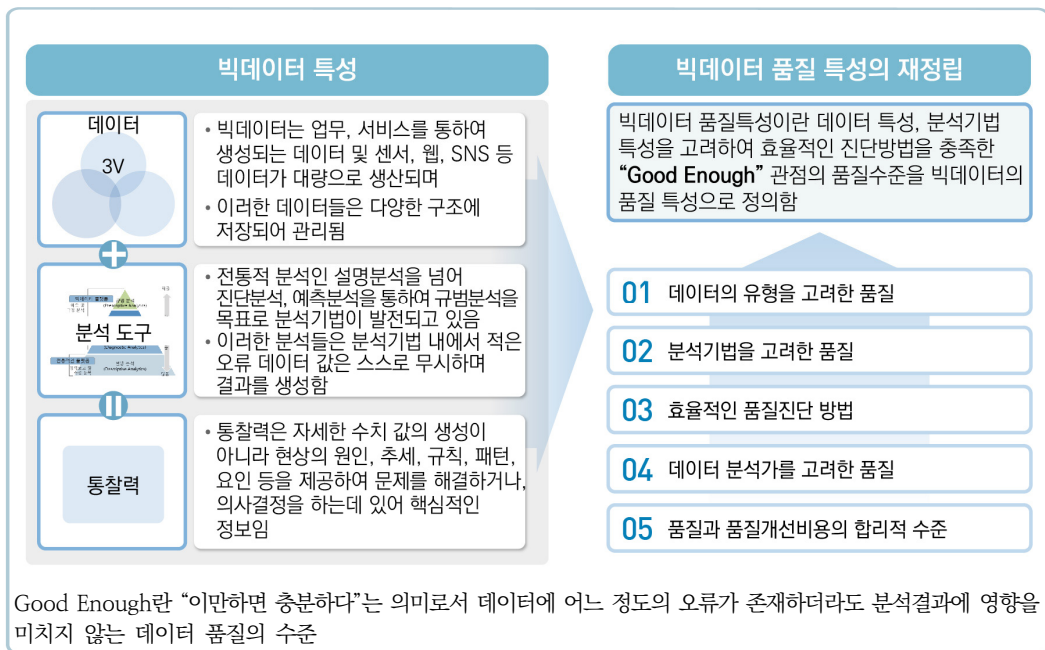
2.1 빅데이터 품질특성

2.1.1 빅데이터 특성 정의

빅데이터는 업무, 서비스를 통하여 생성되는 정보 및 센서 정보, 웹/SNS 등 온라인상에서 발생하는 정보 등 다양한 정보 생산 주체별로 정형, 비정형의 다양한 정보가 대량으로 발생하고 있다.

이렇게 수집된 데이터는 과거와 달리 어떤 현상에 대한 원인, 추세, 규칙, 패턴, 요인 등을 분석하여 현상의 문제를 해결하거나, 의사결정을 지원하는 핵심적인 정보로 제공된다, 즉 통찰력(insight)을 제시하는 것을 빅데이터의 특성으로 볼 수 있다.

그림 6 빅데이터 품질 특성



▶ 빅데이터 품질특성

빅데이터 품질특성이란 데이터 특성, 분석기법 특성을 고려하여 효율적인 진단방법을 충족한 “Good Enough” 관점의 품질 수준을 빅데이터의 품질특성으로 정의한다.

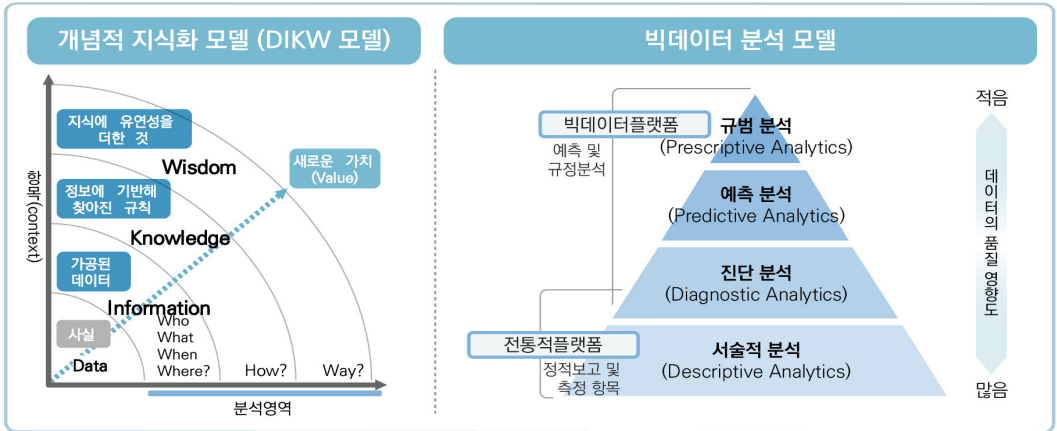
- 데이터의 유형을 고려한 데이터 품질 : 정형, 비정형 및 생산 주체에 따른 품질의 고려
- 분석기술의 특성을 고려한 데이터 품질 : 적은 오류 값은 스스로 무시하며 통찰력을 제시
- 효율적인 품질진단 방법 : 대량 데이터에 대하여 표본조사방법 제시
- 데이터 분석가를 고려한 품질 : 데이터에 대하여 충분한 설명(메타)정보의 제공
- 품질과 품질개선비용의 합리적 수준 : 오류 없는 데이터를 생산하는 비용에 대한 고려

2.1.2 빅데이터 분석모델

빅데이터의 특성인 통찰력(insight)을 찾기 위해서는 정보의 수집뿐만 아니라 이를 분석하는 기술이 중요한 요인으로 자리 잡고 있다.

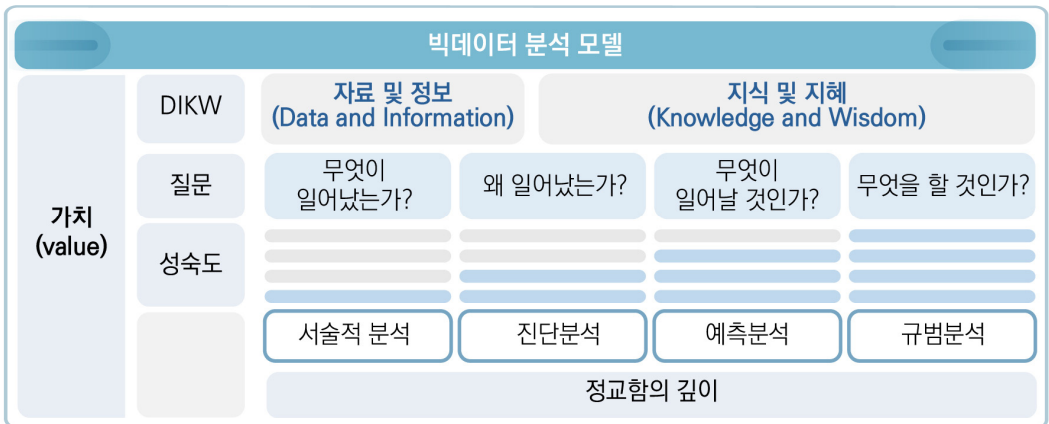
과거 분석기술은 개념적 지식화 모델인 DIKW모델¹⁾에서 “정보(information)”단계까지 진화하였으나 빅데이터의 태동으로 지식화 모델을 이루는 기술이 더욱 정교하게 발전하고 있다.

그림 7 빅데이터 분석 특성



빅데이터의 분석모델은 빅데이터를 기반으로 다양한 분석기법에 따른 분석모델의 발전으로 “진단분석단계”를 넘어 “예측분석단계”가 성장기에 진입하고 있다.

그림 8 빅데이터 분석 유형



1) DIKW 모델은 데이터 피라미드 또는 지식 피라미드라고 불리는 DIKW 피라미드는 데이터(D), 정보(I), 지식(K), 지혜(W)의 상관관계를 설명하는 데 유용하다. 데이터는 관찰이나 조사에서 얻은 사실을 말하는데 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료를 말한다. 정보는 관찰이나 측정을 통해 얻은 데이터를 기초로 의사결정이나 문제해결을 위한 메시지 또는 의미를 담고 있는 지식이나 자료를 뜻한다.

- 서술적 분석(Descriptive Analytics) : 과거에서 현재 데이터를 통해 무엇이 일어났고, 일어나고 있는지를 파악하기 위한 분석으로 특정 시점 또는 특정 기간에 발생한 결과를 보여주는 간단한 보고서 및 시각화를 제공한다.
 - 진단 분석(Diagnostic Analytics) : 과거 데이터를 통해 왜 일어났는지 찾기 위한 분석으로 발생 패턴을 파악하거나, 데이터 분류 또는 원인의 요인을 찾는 분석으로 고급 기능을 통해 분석가는 데이터를 자세히 조사하고 주어진 상황의 근본 원인을 파악한다.
 - 예측 분석(Predictive Analytics) : 현재 생성되는 데이터를 통해 무엇이 일어날 것인지 예측하는 것으로 현재 상태에 대한 확률을 구하여 현상을 예측하는 분석으로 고급 알고리즘인 인공지능과 기계 학습 기술을 사용한다.
 - 규범 분석(Prescriptive Analytics) : 규범 분석은 조직에 원하는 결과를 달성하기 위해 수행해야 할 작업을 알려준다.
- ▶ 빅데이터의 분석모델(Analytics model)²⁾에 적용되는 알고리즘(Algorithm)³⁾들은 적은 입력 오류 값을 스스로 기각하며 결과를 생성하는 특성이 있어 데이터 품질이 결과에 미치는 영향도가 전통적 분석기법보다 상대적으로 적어지는 특성을 보인다.

2.2 데이터 분류체계 정의

빅데이터의 데이터 분류체계는 데이터 유형에 따른 “데이터 유형 분류체계”와 데이터의 저장 구조에 따른 “데이터 구조 분류체계”로 정의한다.

- ▶ 데이터 유형 분류체계는 생산 주체에 따라 “프로세스가 생성”, “기계가 생성”, “사람이 생성”으로 분류하며, 자료출처에 의하여 “업무정보”, “생체정보”, “사물인터넷정보”, “지식정보”, “웹/SNS정보” 등으로 분류 유형 그룹을 정의하여 “메타데이터, 기준데이터, 거래데이터 등”의 10개의 데이터 유형으로 분류하여 데이터 유형 분류체계를 정의한다.
 - ▶ 데이터 구조 분류체계는 컨테츠 형식에 따라 “정형”, “반정형”, “비정형”으로 분류되며, 저장유형에 따라 “RDB”, “No-SQL”, “비구조”로 분류할 수 있으며, 이를 세분화하면 “관계구조, 키-값 구조, 컬럼 패밀리구조 등”의 6개 데이터 구조로 분류하여 데이터 구조분류체계를 정의한다.
- 다양한 유형의 자료를 수집·가공·분석하는 빅데이터 환경에서 자료의 체계적인 분류를 통하여 효율적인 품질진단의 수행을 위하여 데이터에 대한 분류체계를 제시한다.

2) 분석모델은 어떤 현상 또는 프로세스를 해결 가능한 방정식의 집합으로 표현한 것.

3) 알고리즘은 어떤 문제를 해결하기 위해 정해진 일련의 절차나 방법

2.2.1 데이터 유형 분류체계

데이터 유형 분류체계는 데이터의 유형에 따라 아래와 같이 정의한다.

그림 9 빅데이터 유형 분류체계

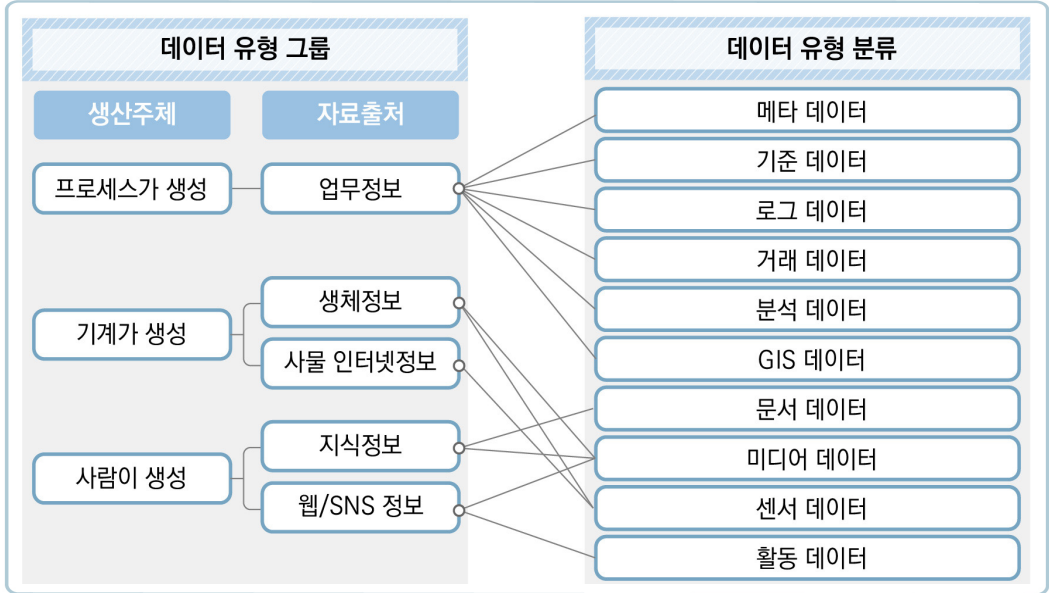


표 3 데이터 유형 분류체계

자료 유형	설명
프로세스가 생성	<ul style="list-style-type: none"> 기관 및 기업의 업무시스템
업무정보	<ul style="list-style-type: none"> 정보시스템 및 서비스로 생성된 데이터
기계가 생성	<ul style="list-style-type: none"> 의료장비, 생산설비, 각종 센서 등 기계장치
생체정보	<ul style="list-style-type: none"> 의료장비, 신분인식기 등으로 생성된 신체에 관한 데이터
사물인터넷정보	<ul style="list-style-type: none"> 기상, 환경, 위치, 가전 등 목적성 장비 및 개인 장치(모바일) 등 센서를 통하여 생성된 데이터
사람이 생성	<ul style="list-style-type: none"> 개인
지식정보	<ul style="list-style-type: none"> 기관의 지식 데이터, 논문, 문서, 미디어 등
웹/SNS정보	<ul style="list-style-type: none"> 온라인 가격정보 온라인 검색, 등록 정보 등 트위터 등 공개 SNS 자료 등

I

II

III

IV

데이터 품질관리 방안

V

표 4 데이터 유형 분류체계 설명

데이터 유형	설명	예
메타 데이터 (Meta data)	<ul style="list-style-type: none"> 데이터에 대한 표준화 정보 및 데이터 특성을 설명하는 요약 정보로 정보명, 주제영역, 품질수준, 다른 데이터와 연관성 등을 관리하는 정보 	<ul style="list-style-type: none"> 표준사전, 데이터셋의 특성정보 (정보명, 위치, 주제영역, 출처, 중요도, 품질수준, 연관성 등)
기준 데이터 (Master data)	<ul style="list-style-type: none"> 업무 프로세스의 중심이 되는 공통 개념의 데이터로 값이 자주 변경되지 않는 데이터 및 참조 데이터(데이터의 값을 참조하는 코드성 데이터)를 포함 	<ul style="list-style-type: none"> 제품정보, 고객정보, 사업자정보 등 지역코드, 성별코드, 학력코드 등
로그 데이터 (Log data)	<ul style="list-style-type: none"> 시스템이 생성한 Log 데이터, 웹상의 모든 흔적의 Raw file 형식 데이터 	<ul style="list-style-type: none"> 로그기록, WebLog(Raw file) 등
거래 데이터 (Transaction data)	<ul style="list-style-type: none"> 기업/기관의 고유한 업무활동 및 서비스로 생성되는 정보 	<ul style="list-style-type: none"> 신용/직불카드 거래 자료 및 금융 거래 자료, 쇼핑자료
분석 데이터 (Analytics data)	<ul style="list-style-type: none"> 집계 데이터 및 분석을 통하여 결과로 생성된 정보 	<ul style="list-style-type: none"> 업종별 매출현황, 서울 인구이동 분포, 종로 상권분석 등
GIS 데이터 (GIS data)	<ul style="list-style-type: none"> GIS 시스템에 의하여 생성한 공간 정보(벡터(Vector), 래스터(Raster) 데이터)와 속성정보로 구성된 정보 	<ul style="list-style-type: none"> 행정구역도, 지하매설물도, 산림도, 정사영상 등
문서 데이터 (Document data)	<ul style="list-style-type: none"> 문서 작성기로 생성한 문서 정보(고유의 저장형식을 가진) 	<ul style="list-style-type: none"> 일반문서, 논문, 보고서 등
미디어 데이터 (Multimedia data)	<ul style="list-style-type: none"> 사진, 영상, 음성 등 미디어 정보 	<ul style="list-style-type: none"> 사진, 위성사진, 항공사진, 영상, 음성 등
센서 데이터 (Sensor data)	<ul style="list-style-type: none"> 추적장치 정보(Tracking device data)를 포함한 각종 센서를 통하여 생성되는 정보 	<ul style="list-style-type: none"> 위치, 기상, 수질, 대기, IoT, 교통 등
활동 데이터 (Online behavior data)	<ul style="list-style-type: none"> 의견정보 (Opinion data), 웹 검색정보 등을 포함한 온라인상에서 생성된 것으로 분석을 위해 가공된 정보 	<ul style="list-style-type: none"> 인터넷 검색 및 페이지뷰 인덱싱 정보 (WebLog) 등 및 이메일, 카카오톡, 트위터 등 공개 자료

2.2.2 데이터 구조 분류체계

데이터 구조 분류체계는 데이터의 저장 구조에 따라 아래와 같이 정의한다.

그림 10 데이터구조 분류체계

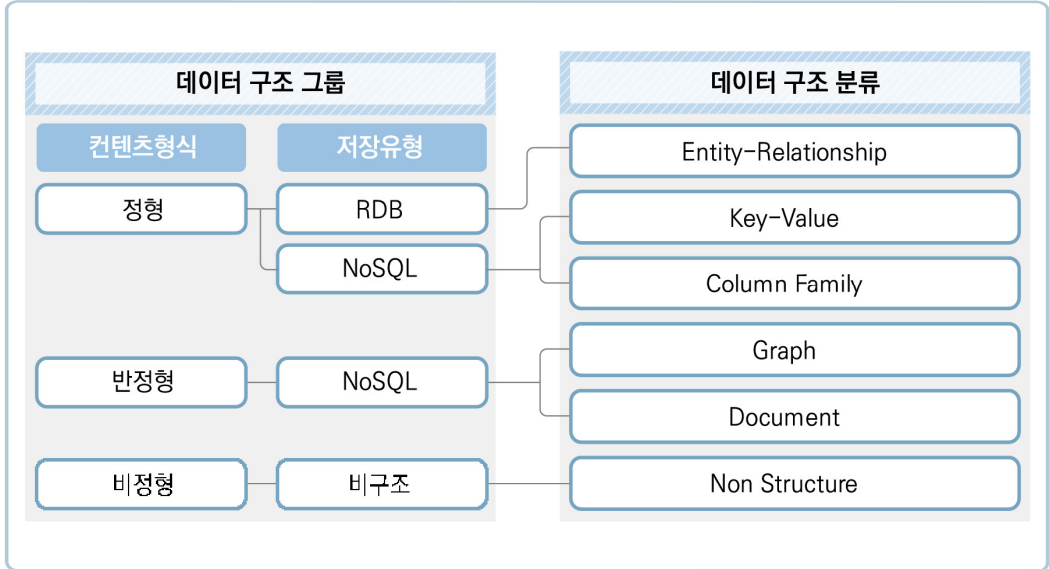


표 5 데이터 구조 분류체계

데이터 구조	설명
RDB	<ul style="list-style-type: none"> 업무시스템에서 사용되는 데이터베이스
관계구조	<ul style="list-style-type: none"> 객체간의 관계가 정의된 구조
NoSQL	<ul style="list-style-type: none"> 의료장비, 생산설비, 각종 센서 등 기계장치
키-값 구조	<ul style="list-style-type: none"> Key와 Value로 구성된 구조
컬럼 집합 구조	<ul style="list-style-type: none"> 하나의 Key에 여러 Column 속성으로 구성된 구조
그래프 구조	<ul style="list-style-type: none"> 정보 간의 연결 관계를 맺는 구조
문서 구조	<ul style="list-style-type: none"> JSON, XML과 같이 값과 구조가 같이 정의된 구조
비 구조	<ul style="list-style-type: none"> 파일 시스템에 직접 저장되는 구조
비 구조	<ul style="list-style-type: none"> 정보가 집합체의 성격을 갖거나 정보를 세분화하여 정형, 반정형 구조에 정보를 저장할 수 없어 파일 시스템에 저장되는 구조

03

데이터 품질진단 방안



3.1 품질진단 대상 선정

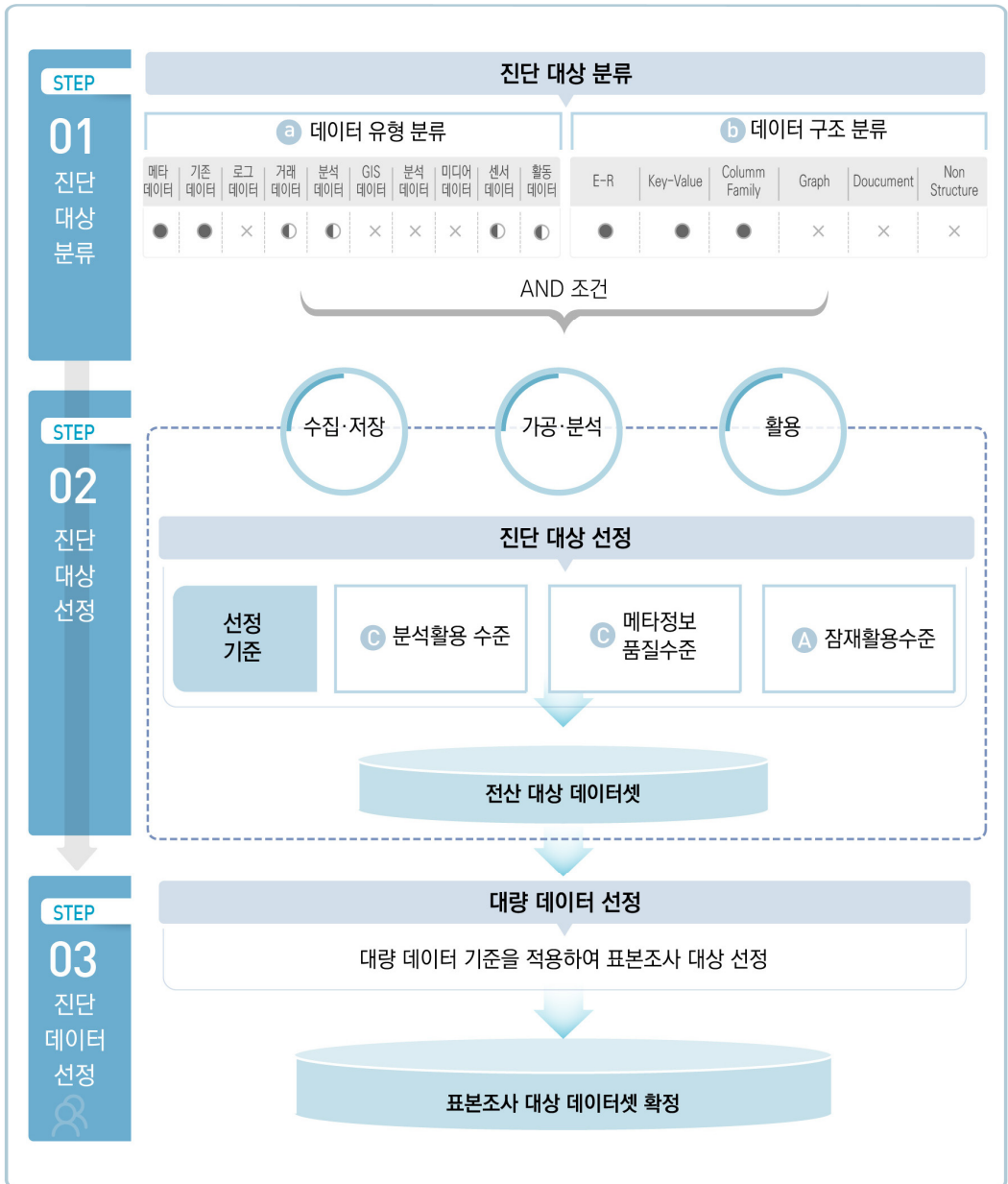
빅데이터의 품질진단을 효율적으로 수행하기 위하여 데이터 분류체계에 의한 진단대상을 분류(㉠,㉡)하고, MECA 기법에 따라 1) 필수진단 대상을 우선 선정하며, 데이터 중요도에 따라 2) 선택적 진단 대상을 선정하며, 데이터 품질이슈에 따라 3) 이슈별 추가 진단대상을 선정한다.

표 6 MECA 기법(대상 선정 기준)

용어	정의	설명
Mandatory	필수 진단대상	<ul style="list-style-type: none"> 메타데이터, 기준데이터 등 핵심 데이터로서 품질평가를 위한 필수 데이터
Exceptional	진단 제외 대상	<ul style="list-style-type: none"> 시스템로그, 임시테이블, 백업테이블 등 품질평가가 불필요한 데이터 및 측정 도구가 없는 비정형 데이터
Conditional	중요도 평가에 의한 진단대상	<ul style="list-style-type: none"> 분석활용수준, 메타품질수준 등을 평가하여 선택적으로 품질평가를 위한 데이터
Additional	이슈별 추가진단대상	<ul style="list-style-type: none"> 데이터 품질이슈에 따라 정의한 추가 선정기준을 충족하는 데이터셋

분류기준에 따라 선정된 데이터 중에 대량 데이터를 관리하는 레코드 건수가 많은 테이블에 대하여 통계적 기법을 적용하여 표본조사에 의한 품질진단 대상으로 선정한다.

그림 11 데이터셋(테이블) 진단대상 선정절차



▶ 진단대상 분류

- 품질진단 대상 데이터셋(테이블) 선정을 위하여 데이터 유형분류와 데이터 구조분류를 이용하여 진단 대상 분류기준 및 매트릭스를 정의하여 수집된 데이터를 체계적으로 분류한다.
- 진단 대상 분류기준 및 매트릭스에 따른 진단 대상 분류는 플랫폼 자체적으로 마련할 수 있다. 단, 가이드에서 제시하는 분류와 1:1 또는 1:N(상세분류)으로 매핑 되어야 한다.

그림 12 진단 대상 분류

분류 기준 정의

① 유형분류기준

- ④ 필수 품질수준평가를 위한 데이터 후보
- ⑤ 분석에 이용되지 않는 데이터 유형은 제외
- ⑥ 특정도구가 필요한 데이터 유형은 제외
- ⑦ 정량적 측정이 어려운 데이터 유형은 제외
- ⑧ 특정 조건을 충족하는 경우 선택적인 대상

② 구조분류기준

- ⑤ 데이터 구조분류 중 진단도구의 부재로 진단이 불가능한 데이터 구조는 제외함

범
례

- ④ 필수진단 대상
- ⑤ 진단제외 대상
- ⑧ 선택진단 대상
- 정성적조사대상

분류기준		데이터 구조분류					
		RDB E-R	Document				Non Structure
			Key-Value	Column Family	Graph	Document	
데이터 유형 분류	메타 데이터	●	●	●	×	×	×
	기준 데이터	●	●	●	×	×	×
	로그 데이터	×	×	×	×	×	×
	거래 데이터	④	④	④	×	×	×
	분석 데이터	④	④	④	×	×	×
	GIS 데이터	④	×	×	×	×	×
	문서 데이터	○	○	○	×	×	○
	미디어 데이터	○	○	○	×	×	○
	센서 데이터	④	④	④	×	×	×
	활동 데이터	④	④	④	×	×	×

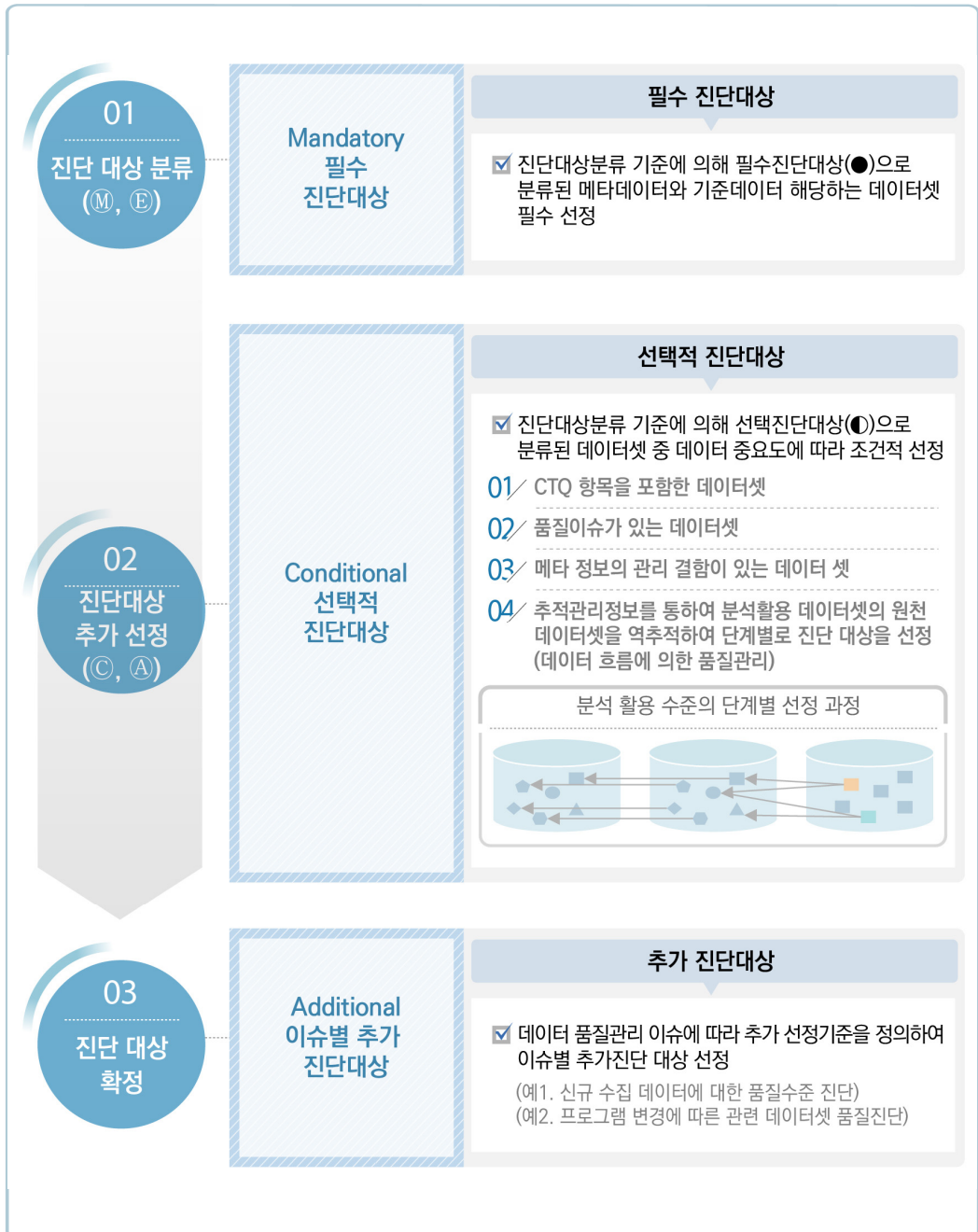
표 7 진단 대상 분류 설명

항 목	내 용	비고
● ④필수 진단 대상	E-R, Key-Value, Column Family 구조로 저장 되어 관리되는 데이터 중 품질관리의 기준 데이터로 사용되는 메타데이터 및 기준데이터는 진단 대상으로 분류	
④ ⑧선택 진단 대상	데이터 유형이 거래데이터, 분석 데이터, GIS 데이터, 센서 데이터, 활동데이터에 대하여 선택 진단 대상으로 분류	• 진단 대상선정 기준(⑧)에 의해 추가 선정하여 진단 대상을 확정함 • GIS는 GIS엔진기반 진단 도구 필요
X ⑤진단 제외 대상	데이터 유형에 관계없이 그래프, 문서구조(JSON, XML)로 저장된 데이터는 진단 대상에서 제외하며, 분석과 관련이 낮은 기록데이터(Log data)도 진단 대상에서 제외 대상으로 분류	• 단, E-R, Key-Value, Column Family로 변환 시 진단 가능
○ 정성적 조사 대상	비구조의 성격을 가지는 문서, 미디어(이미지, 영상, 음성) 유형의 데이터는 개별적인 품질진단은 생략하고 조사표(진단체크리스트)를 이용하여 포괄적인 품질의 수준을 정성적으로 측정 대상으로 분류	• 단, 비정형 데이터에 대한 메타정보는 진단 함(완전성)

▶ 진단 대상 선정

진단 대상 분류기준에 의해 필수진단대상을 선정하고, 선택 진단대상 중 선정기준에(분석활용수준, 메타정보품질수준, 잠재 활용수준)따라 평가 후 추가 선정하여 품질진단 대상을 확정한다.

그림 13 진단 대상 및 선정 기준



I

II

III

IV

데이터 품질진단 방법

V

표 8 진단 대상 선정 기준 설명

구분		정의	기준
Mandatory 필수 진단대상	메타데이터 기준데이터	<ul style="list-style-type: none"> 메타 데이터와 기준데이터에 해당하는 데이터 셋 선정 	<ul style="list-style-type: none"> 데이터 유형 분류에 의한 메타 데이터와 기준데이터
Conditional 선택적 진단대상	CTQ 대상	<ul style="list-style-type: none"> 정의되어 관리되고 있는 CTQ에 해당하는 항목(컬럼)을 포함하고 있는 데이터셋 	<ul style="list-style-type: none"> CTQ 항목을 포함한 데이터셋
	분석활용 대상	<ul style="list-style-type: none"> 데이터의 활용도를 산정하여 활용도가 높은 데이터셋 <ul style="list-style-type: none"> - 분석·활용에 이용된 원천 데이터셋을 단계별로 역추적한 데이터셋의 이용 수 - 무/유상 제공 데이터셋의 6유통횟수 	<ul style="list-style-type: none"> 유통(분석인용/유통)된 데이터셋
	품질이슈 대상	<ul style="list-style-type: none"> 품질이슈가 있는 데이터셋 	<ul style="list-style-type: none"> 품질이슈가 있는 데이터셋
	메타정보 누락, 오류 대상	<ul style="list-style-type: none"> 데이터셋의 메타정보에 대하여 품질 수준을 측정하여 품질이슈가 있는 데이터셋 <ul style="list-style-type: none"> - 데이터셋 표준화 메타정보 누락, 오류 - 데이터셋에 대한 설명 메타 누락, 오류 	<ul style="list-style-type: none"> 메타정보 누락, 오류가 있는 모든 데이터셋 <ul style="list-style-type: none"> - 표준화 메타 누락, 오류 - 설명메타 누락, 오류
Additional 이슈별 추가 진단대상	이슈별 추가진단 대상	<ul style="list-style-type: none"> 데이터 품질관리 이슈에 따라 정의한 선정을 충족하는 데이터셋 <ul style="list-style-type: none"> - 예1. 신규 수집 데이터에 대한 품질수준 진단 - 예2. 프로그램 변경에 따른 관련 데이터셋 품질진단 	<ul style="list-style-type: none"> 이슈에 따라 선정기준 정의 <ul style="list-style-type: none"> - 데이터 품질이슈에 따라 정의한 추가 선정기준을 충족하는 데이터셋

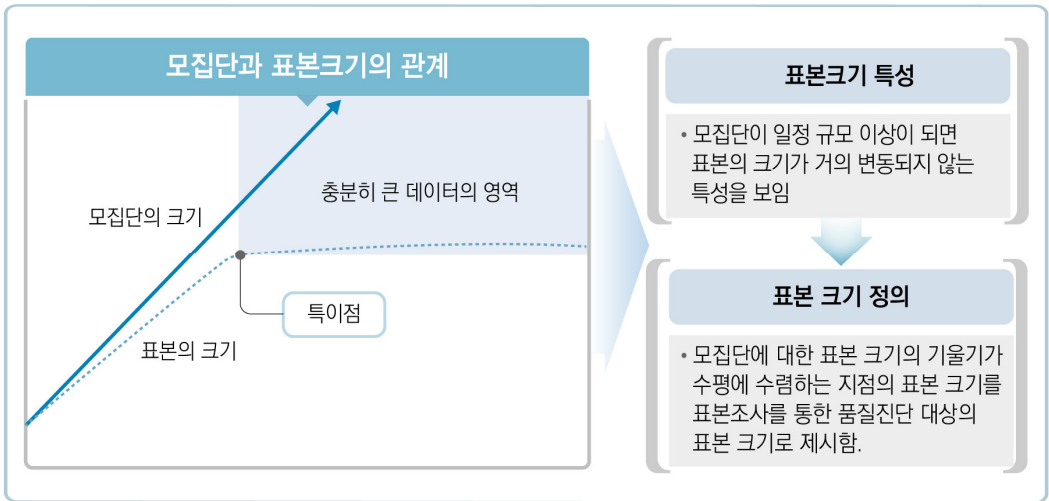
3.2 대량 데이터 선정

빅데이터는 대량으로 데이터가 발생하는 특성을 보여 데이터 품질관리를 효율적인 수행을 확보하기 위하여 대량 데이터셋을 선정하여 표본추출을 통한 품질진단을 제시하고자 한다.

3.2.1 대량 데이터의 선정기준

- ▶ 대량 데이터(Big table)의 효율적 품질진단 방법으로 통계적 방법에 따른 표본조사 방법을 제시하며, 대량 데이터의 선정기준으로 모집단 규모에 따른 표본 크기의 변화 특성을 분석하여 대량 데이터의 표본 크기를 정의한다.
- ▶ 대량 데이터의 기준은 모집단의 규모가 증가하더라도 일정 규모부터 표본의 크기가 미미하게 증가하는 현상을 보여 이 시점의 모집단의 크기를 대량 데이터로 정의할 수 있다.
- ▶ 표본의 크기가 미미하게 증가하는 데이터의 규모는 약 3천만 또는 4천만 정도로 볼 수 있다. 따라서 대량 데이터의 양적 크기를 약 3천만 건으로 정의할 수 있다.

그림 14 모집단과 표본 크기의 관계



- ▶ 표본의 크기를 결정하는 표본설계변수는 모집단의 규모, 신뢰수준, 표본오차에 의하여 표본크기가 결정된다. 일반적으로 표본조사의 신뢰도를 결정하는 신뢰수준은 95%를 적용하고 있으나 품질진단의 신뢰도를 향상하기 위하여 신뢰수준을 99%로 정하며, 표본오차의 크기는 0.1%로 정하여 표본의 크기를 산출한다.

표 9 신뢰수준 및 표본오차에 따른 표본 크기의 비교

구분	신뢰수준					
	99%			95%		
	표본오차			표본오차		
	0.1%	0.05%	0.03%	0.1%	0.05%	0.03%
대량 데이터 규모	3,000만	3,000만	3,000만	3,000만	3,000만	3,000만
표본 크기	157만	540만	1,144만	93만	340만	787만

- ▶ 따라서 표본설계기준의 신뢰수준은 99%, 표본오차는 0.1%로 정의한다.
- ▶ 또한, 대량 데이터의 규모는 표9에 의하여 모집단의 규모가 1천만일 경우 표본의 크기는 142.6만, 3천만일 경우 표본 크기는 157.6만, 1억일 경우 표본의 크기는 163.6만으로 모집단 1천만과 3천만의 표본의 크기 차이는 약 15만, 모집단 3천만과 1억의 표본 크기 차이는 5만9천으로 모집단이 3천만부터 표본 크기의 변화가 미미하게 나타나 대량 테이블의 기준을 3천만으로 제시한다.
- ▶ 따라서 대량 데이터의 표본 크기는 대량 데이터 크기 기준인 모집단이 3천만일 때 표본 크기인 157.6만을 표본조사에 필요한 표본 크기 기준으로 정의한다.

표 10 모집단 규모별 표본 크기

구분	내용						
표본추출기준	신뢰수준 99%, 표본오차 0.1%						
데이터 규모	무한대	100억	50억	10억	1억	3천만	1천만
표본 크기	166.4만	166.3만	166.3만	166.1만	163.6만	157.6만	142.6만

구분	의미	설명
신뢰수준	<ul style="list-style-type: none">신뢰수준 99%는 해당 표본조사를 99% 믿을 수 있다는 뜻이 아니라 같은 조사를 100번 하면 오차범위 내 동일한 결과가 나올 가능성이 99번이라는 의미이다.	<ul style="list-style-type: none">제품의 품질조사에서 99% 신뢰수준이라고 함은, 100번 조사를 하였을 때 99번은 같은 결과가 나온다는 것으로, 이는 표본을 골고루 잘 선택하였다는 뜻이다.
표본오차	<ul style="list-style-type: none">표본으로부터 계산되는 추정치와 모집단의 값과 차를 표본오차라 한다.	<ul style="list-style-type: none">품질검사를 하여 불량품이 없을 때를 0, 전부가 불량품일 때를 100으로 했을 때, 만약 조사한 불량률이 3%이고, 표본오차 $\pm 0.1\%$라고 한다면 불량률은 2.9% ~ 3.1%의 범위에 존재한다.

➤ 그러나 표본조사를 통한 품질진단을 위한 모집단의 규모가 일정하지 않으므로 모집단 규모에 따른 표본이 과소 추출되는 것을 방지하기 위하여 모집단의 규모가 무한대일 때 표본 크기인 166.4만 보다 큰 167만을 표본의 크기로 제시한다.

표본크기의 정의

- 표본설계기준은 신뢰구간은 99%, 표본오차는 0.1%로 제시한다.
- 대량 데이터의 규모 기준은 3천만으로 제시한다.
- 대량 데이터의 규모 기준 3천만의 표본크기인 157.6만을 표본조사에 필요한 표본 크기의 기준으로 정의한다.
- 표본조사를 통한 품질진단을 위한 표본 크기는 모집단의 규모와 관계없이 **표본 크기를 167만으로 제시한다.**

3.2.2 대량 데이터의 품질진단 방안

표본조사를 통한 품질진단의 수행은 보유한 데이터의 품질 수준을 측정하기 위하여 이용되며, 데이터의 품질이 일정수준 이하로 조사되거나 일정수준 이상의 품질로 조사되더라도 데이터의 품질 개선 수행이 요구되면 오류 데이터의 확인을 위하여 전수조사를 수행하여야 한다.

그림 15 대량데이터의 품질진단 방안



표본 진단 활용

- 표본조사에 통한 품질진단은 대량의 데이터를 관리하는 Big Table의 품질 진단의 비용(시간, 자원)의 비효율성을 위한 방안으로 제시한다.
- 표본조사에 따른 품질진단 적용 여부는 기관이 자율적으로 결정한다.
- 표본조사에 통한 품질진단은 신속한 품질 수준의 평가를 위한 진단 수행 시 적용하는 것을 권장하며 개선을 위한 진단 시에는 반드시 전수조사로 품질진단을 수행한다.

3.3 빅데이터 품질 기준

▶ 빅데이터 품질 기준 개요

품질 기준은 6-시그마의 확률값과 정부의 공공데이터 수준 평가 기준 및 한국데이터베이스진흥원의 K-DATA 데이터인증 기준을 바탕으로 빅데이터의 특성을 반영한 “Good Enough” 관점의 품질 기준의 마련을 통하여 빅데이터의 데이터 품질 기준을 제시한다.

그림 16 빅데이터 품질기준 개요

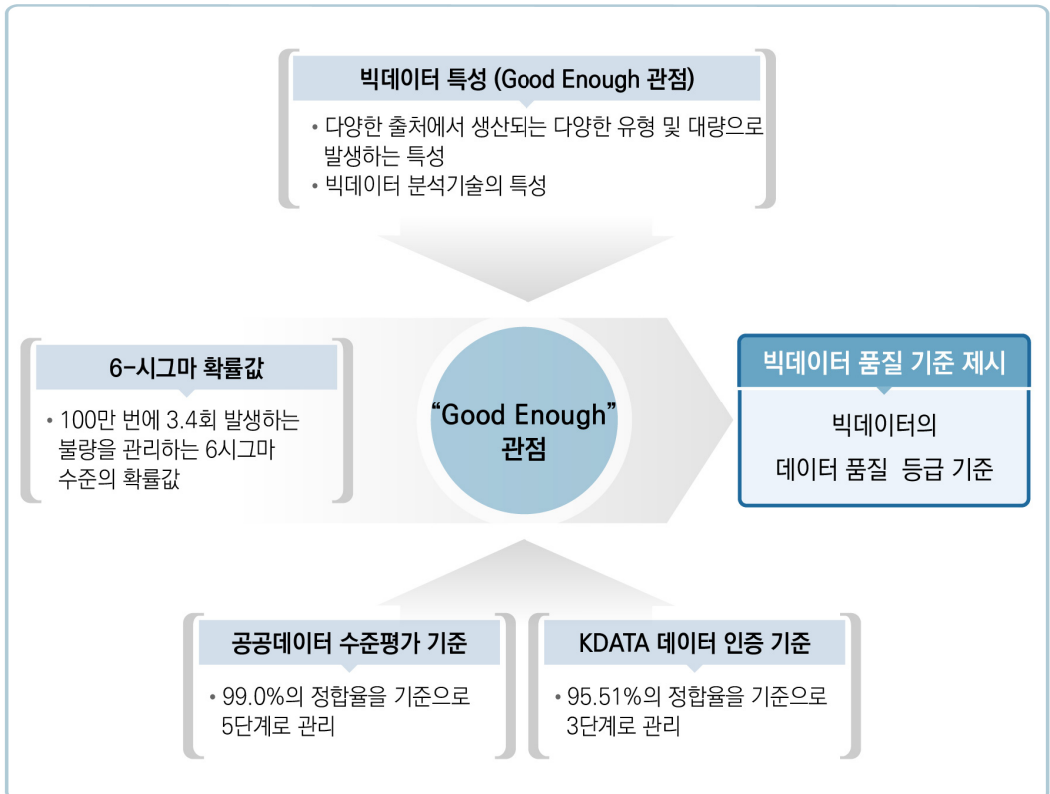
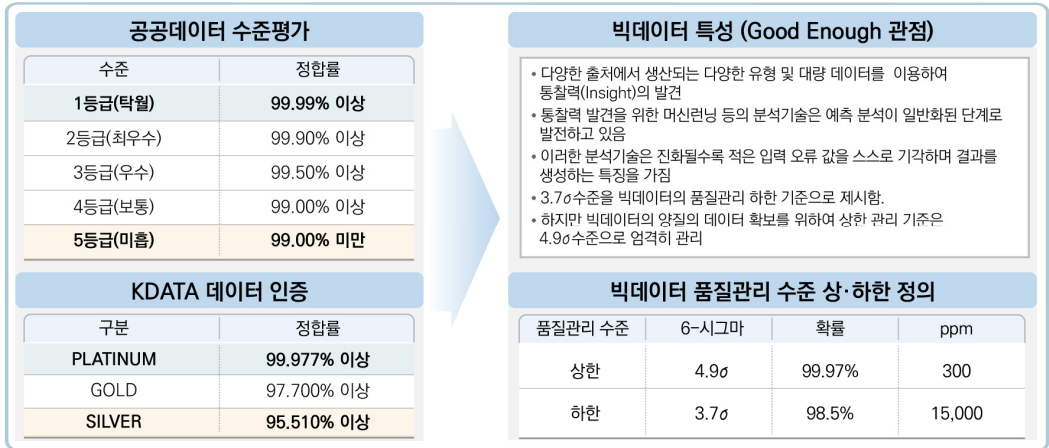


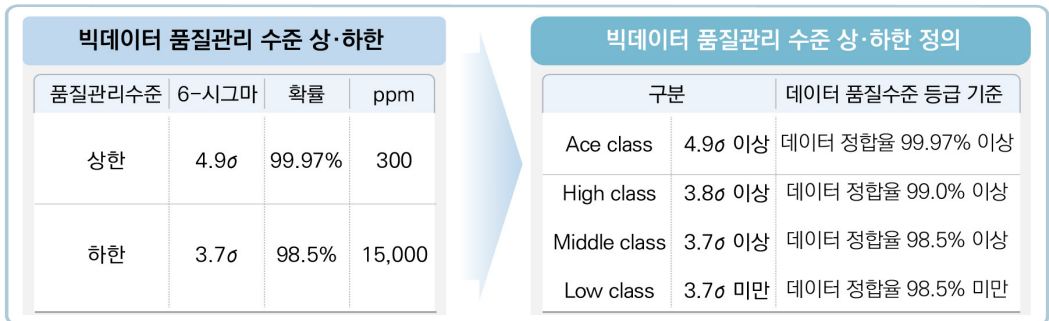
그림 18 빅데이터 품질의 6-시그마 품질관리 수준 정의



▶ 빅데이터 품질관리 수준 정의

표준정규분포의 분포를 이용한 6-시그마(품질관리방법론)의 개념을 적용한 KDATA의 데이터 품질 인증 등급 기준(3단계), 공공데이터 수준평가 기준(5단계)을 바탕으로 빅데이터 품질의 6-시그마 수준의 상·하한 관리 수준을 정의한다.

그림 19 빅데이터 품질수준 등급



▶ 빅데이터 품질수준 등급

빅데이터의 6-시그마 수준 상·하한 (4.9σ~3.7σ) 기준 범위에서 빅데이터의 데이터 품질수준 등급을 4단계로 정의함.

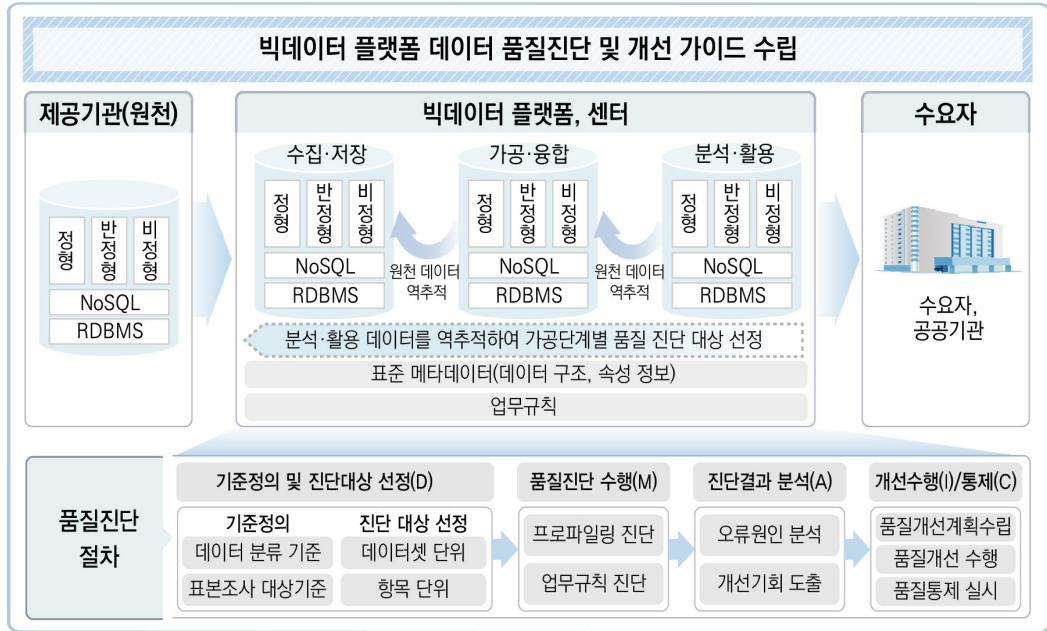
표 11 빅데이터 품질수준 등급

구분		데이터 품질수준 등급 기준
Ace class	4.9 σ	데이터 정합률 99.97% 이상
High class	3.8 σ	데이터 정합률 99.0% 이상
Middle class	3.7 σ	데이터 정합률 98.5% 이상
Low class	3.7 σ	데이터 정합률 98.5% 미만

3.4 데이터 품질진단 목표모델

빅데이터 플랫폼 데이터 수집/저장, 가공/융합, 분석/활용 단계에 따라 빅데이터 특성을 고려한 품질 진단 목표모델을 제시한다.

그림 20 빅데이터 품질진단 목표 모델



3.5 데이터 품질관리를 위한 전제조건

빅데이터 플랫폼의 데이터 품질진단을 위한 전제조건으로 데이터 분류체계, 및 데이터표준(메타 데이터, 표준코드 등), 품질진단 도구의 도입 및 진단환경 구축이 요구된다.

표 12 데이터 품질관리를 위한 전제조건

전제조건	내용
품질진단 조직 운영 및 R&R 정의	• 품질진단 수행을 위한 조직 및 담당자를 지정하고 품질진단을 위한 역할 및 책임을 정의
빅데이터 플랫폼에 적합한 데이터 분류체계 마련	• 빅데이터 플랫폼의 다양한 구조 및 유형의 데이터 품질을 관리하기 위한 적합한 데이터 분류체계 마련 (*데이터분류체계는 플랫폼 자체적으로 마련할 수 있다. 단, 가이드에서 제시하는 분류와 1:1 또는 1:N으로 매핑 되어야 함)
빅데이터 특성을 고려한 메타데이터 및 표준코드 필요	• 데이터 품질진단 시 참조하는 표준정보(표준용어, 표준도메인, 표준코드 등) 및 메타 데이터 필요
효율적인 품질진단 수행을 위한 품질진단 도구 및 진단환경 구축	• 효율적인 품질진단 수행을 위한 품질진단 도구 및 진단환경 구축

▶ 품질진단을 위한 메타 데이터 정의

빅데이터의 품질특성을 고려한 대량데이터의 효율적 품질진단을 위한 필수 메타 데이터를 정의는 표13과 같다.

표 13 품질진단을 위한 메타 데이터 정의

구분	메타 항목	설명
분류 체계 관리 항목	유형분류명	<ul style="list-style-type: none"> 데이터 유형 분류체계에 따른 데이터셋의 유형 분류 (메타, 기준, 거래, 분석, GIS 데이터 등) 데이터셋(테이블) 단위의 메타데이터 정보 "2.2.1 데이터 유형 분류체계" 참조
	구조분류명	<ul style="list-style-type: none"> 데이터 구조 분류체계에 따른 데이터셋의 저장 DBMS 유형 분류 (RDB, Key-Value, Column-Family, Graph 등) DB 단위의 메타데이터 정보 "2.2.2 데이터 구조 분류체계" 참조
분석 활용 관리 항목	원천 데이터 정보	<ul style="list-style-type: none"> 수집, 가공·분석, 활용 등 각 단계에 사용된 원천 데이터 정보 (데이터셋명 (테이블), 항목명(컬럼)) 항목(컬럼) 단위의 메타데이터 정보
	적재 주기	<ul style="list-style-type: none"> 연속적인 데이터의 적재를 요하는 데이터셋의 데이터 적재 주기 (예: 1회성, 매월, 분기, 반기 등) 데이터셋(테이블) 단위의 메타데이터 정보

I

II

III

IV

데이터 품질진단 방법론

V

04

품질진단 및 개선 가이드구성



「품질진단 및 개선 가이드」내용은 6-시그마 방법론(DMAIC) 활동단계를 기반으로 품질진단 대상 정의, 품질진단 수행, 진단결과 분석, 개선 수행, 품질 통제 순으로 작성되었다.

‘DMAIC’란, Define, Measure, Analyze, Improve, Control의 줄임말로 총 16개의 각 단계별 세부 절차를 포함하고 있다.

그림 21 품질진단 및 개선 절차



▶ [D.품질진단 대상 정의]

단계에서는 데이터 전체를 진단대상으로 하지 않고, 핵심품질 관리대상을 선정하여 품질진단을 수행한다. 대량 데이터의 경우에는 데이터 규모에 따라 조사방법(전수조사, 표본조사)을 선택하여 수행한다.

- 품질진단 및 개선에 대한 계획서를 작성하고, 품질진단 수행 대상이 되는 데이터셋(테이블) 및 항목(컬럼)을 선정
- 정량적인 측정이 가능하고 분석 활용도가 높으며 품질이슈가 있거나 지속적인 품질관리가 필요한 데이터를 선정

▶ [M.품질진단 수행]

단계에서는 데이터의 품질 수준을 파악하고 시스템에서 발생하는 잠정 오류를 도출한다.

- 품질진단 목적, 진단 범위 및 일정 등을 고려하여 품질진단 계획을 수립
- 진단기법(프로파일링, 업무규칙, 체크리스트)을 적용하여 품질진단규칙을 정의하고, 정의한 품질진단규칙을 기준으로 품질진단을 실시
- 품질진단 대상에 대해 잠정적 오류를 도출하여 데이터 품질 수준을 파악

▶ [A.진단결과 분석]

단계에서는 진단 결과 데이터 오류에 대한 근본 원인을 분석하고, 오류 유형별 개선 기회를 도출한다.

- 진단결과 오류에 대해서 오류 유형을 구분하고 원인을 분석
- 오류 발생의 근본 원인을 제거하기 위하여 전체 시스템 운영 관점에서 오류 개선 기회를 도출

▶ [I.개선 수행]

단계에서는 품질개선 과제 정의 및 품질개선 계획을 수립하고, 개선 환경을 준비하여 실제 개선을 수행한다.

- 개선 오류를 유발하는 근본 원인을 제거하기 위하여 품질개선 과제를 정의하고, 수립한 계획에 따라 개선을 수행
- 개선 효과가 크고 중요한 문제들을 우선적으로 선정하여 개선

▶ [C.품질 통제]

단계에서는 개선 사업의 효과에 대한 결과를 평가하고, 고품질의 데이터를 확보하기 위한 통제활동을 수행한다.

- 개선결과 평가 후, 품질유지 활동과 개선 활동이 지속적으로 실행될 수 있도록 체계적인 품질 목표를 수립
- 고품질의 데이터를 확보하기 위해, 지속적인 교육 및 모니터링 등의 품질통제 활동을 실시

01

개요



정의

“품질진단 대상 정의(Define)”은 품질진단을 수행 할 대상이 되는 데이터 셋(테이블) 및 항목(컬럼)을 선정하는 것이다.

정량적인 측정이 가능하고 분석활용도가 높으며 품질이슈가 있거나 지속적인 품질관리가 필요한 데이터를 선정한다.

목적

빅데이터는 다양한 형식의 대량 데이터를 수집하여 분석을 통해 통찰력을 발견하는 특성이 있어 이를 반영한 효율적인 일상적 품질진단 수행을 위해 데이터 전체를 진단대상으로 하지 않고 핵심 품질관리대상을 선정하여 품질진단을 수행하며, 대량 데이터의 경우에는 데이터의 규모에 따라 조사방법(전수조사, 표본조사)을 선택하여 수행 할 수 있다.

절차

품질진단 및 개선 전반에 대한 계획서를 작성하고, 데이터 분류체계(데이터 유형, 데이터 구조) 기준으로 분석활용도가 높은 데이터를 데이터셋(테이블) 단위로 대상을 선정 한 후, 선정 된 데이터셋(테이블)에서 CTQ 항목 또는 분석활용도가 높은 항목(컬럼)을 선정한다.

그림 22 품질진단 대상 정의(Define) 절차



표 14 품질진단 정의(Define) 절차 설명

절 차	내 용
D0. 품질관리 계획	<ul style="list-style-type: none"> 품질진단 및 개선 전반에 대한 계획서를 작성한다 - 산출물 : [서식-1] 품질관리 계획서
D1. 데이터셋(테이블) 단위 선정	<ul style="list-style-type: none"> MECA기법⁴⁾에 근거하여 데이터 분류체계(④데이터 유형, ⑤데이터 구조)를 기준으로 데이터셋의 유형을 분류한다. ③분석활용 관점에서 중요도가 높고, 메타정보 품질수준이 낮고, 신규추가 데이터 중 잠재활용 수준이 높은 데이터셋을 선정한다. 단, 정기진단 시는 전체 테이블을 대상으로 한다 ④대량데이터 기준을 적용하여 조사방식(전수, 표본)을 결정한다. 단, 진단은 표본조사 가능하나 개선은 전체 데이터를 대상으로 한다. - 산출물 : [서식-2] 품질진단대상목록(데이터셋-테이블)
D2. 항목(컬럼) 단위 선정	<ul style="list-style-type: none"> 선정 된 대상 데이터셋(테이블)에서 품질진단 대상 도메인⁵⁾에 해당하는 항목(컬럼)을 선정한다. - 산출물 : [서식-3] 품질진단대상목록(항목-컬럼)

4) 본 가이드에서 제시한 빅데이터 분류체계 및 대량 데이터 품질진단 기반의 단계적 진단대상 선정 기법 (Mandatory, Exceptional, Conditional, Additional의 약자)

5) 품질진단 대상 도메인 : 코드, 여부, 번호, 날짜, 율, 수량, 금액 도메인과 이외 본질식별자로 사용되는 명 도메인 등을 말함.

02

품질관리 계획



품질관리 계획은 데이터 품질진단과 개선 전반에 대한 계획을 수립하는 단계로 품질진단 및 개선의 목적, 데이터베이스 현황, 품질관리조직, 품질진단 및 개선 계획 등을 정의한다.

표 15 품질진단 계획서 구성 항목

항 목	내 용
품질진단 및 개선 목적	<ul style="list-style-type: none"> 품질진단 및 개선 계획에 대한 목적을 기술한다. 예) 품질진단 정기 선행 점검에 따른 품질진단 수행 및 오류 데이터 개선을 수행하고자 함
품질관리 조직	<ul style="list-style-type: none"> 품질관리 수행 조직에 대한 조직도/역할 내역을 기술한다. 예) 수행주체 별 역할 (품질관리 담당자/시스템 품질진단 업무)의 내용
품질진단 대상	<ul style="list-style-type: none"> 데이터베이스 현황에 대해 기술한다. - 저장영역, DBMS, DATA SIZE(GB), 전체 테이블 수 정보 내용
품질개선 계획	<ul style="list-style-type: none"> 품질개선 범위 및 일정을 기술한다. <ol style="list-style-type: none"> 품질진단 범위 예) 품질진단 정기 선행 점검에 따른 품질진단 수행결과에 따라 품질개선 범위 확정 예정 품질진단 수행 일정 - 항목 : 수행TASK, 수행일정, 수행내용

(산출물 : [서식-1]품질관리 계획서)

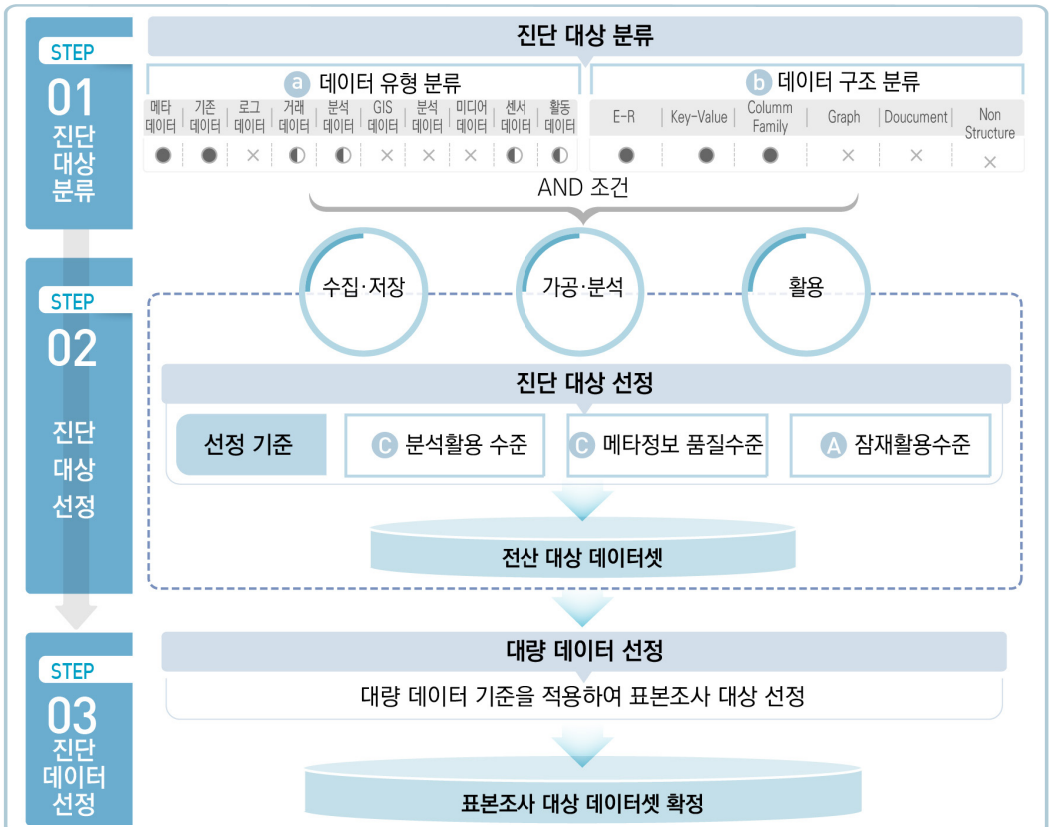
03

데이터셋(테이블) 단위 선정

▶ 데이터셋(테이블) 선정 개요

빅데이터의 일상적 품질진단을 효율적으로 수행하기 위하여 데이터 분류체계에 의한 데이터유형을 분류하고, MECA 기법을 적용하여 1) 필수진단 대상, 2) 데이터 중요도에 따라 선택적 진단 대상, 3) 데이터 품질이슈에 따라 이슈별 추가진단 대상을 선정하고 대량 데이터를 관리하는 빅데이터에 대하여는 통계 기법에 근거하여 표본진단 대상을 선정한다.

그림 23 데이터셋(테이블) 진단대상 선정절차



I

II

III

IV

데이터 품질진단 방법

V

▶ 진단 대상 분류

품질진단 대상 데이터셋(테이블) 선정을 위하여 진단대상 분류(유형분류, 구조분류)를 기준으로 전체 데이터 셋(테이블)에 대해 데이터 유형을 정의한다.

표 16 데이터 유형분류

데이터 유형	설 명	예
메타 데이터 (Meta data)	<ul style="list-style-type: none"> 데이터에 대한 표준화 정보 및 데이터 특성을 설명하는 요약 정보로 정보명, 주제영역, 위치, 중요도, 품질수준, 다른 데이터와 연관성 등을 관리하는 정보 	<ul style="list-style-type: none"> 데이터셋의 특성정보(정보명, 위치, 주제영역, 출처, 중요도, 품질수준, 연관성 등)
기준 데이터 (Master data)	<ul style="list-style-type: none"> 업무 프로세스의 중심이 되는 공통 개념의 데이터로 정보값이 자주 변하지 않는 정보 및 참조정보(Reference data : 데이터의 값을 참조하는 코드성 정보)를 포함 	<ul style="list-style-type: none"> 제품명세, 고객정보, 사업자정보 등 지역코드, 성별코드, 학력코드 등
거래 데이터 (Transaction data)	<ul style="list-style-type: none"> 기업이나 기관의 고유한 업무활동 및 서비스에 의하여 생성되는 정보 	<ul style="list-style-type: none"> 신용/직불카드 및 금융 거래 자료, 유통업체 자료
분석 데이터 (Analytics data)	<ul style="list-style-type: none"> 집계 데이터 및 분석을 통하여 결과로 생성된 정보 	<ul style="list-style-type: none"> 지역별/업종별 매출현황, 서울 인구 이동 분포, 종로 상권분석 등
로그 데이터 (Log data)	<ul style="list-style-type: none"> 시스템이 생성한 Log 데이터, 웹 상의 모든 흔적들의 Raw file 형식 데이터 	<ul style="list-style-type: none"> 로그기록, WebLog (Raw file) 등
GIS 데이터 (GIS data)	<ul style="list-style-type: none"> GIS 시스템에 의하여 생성한 공간 정보로 벡터정보와 속성정보로 구성된 정보 	<ul style="list-style-type: none"> 행정구역도, 지하매설물도, 농경지 지도, 산림도 등
문서 데이터 (Document data)	<ul style="list-style-type: none"> 문서작성기로 생성한 문서 정보(고유의 저장 형식을 가짐) 	<ul style="list-style-type: none"> 일반문서, 논문, 보고서 등
미디어 데이터 (Multimedia data)	<ul style="list-style-type: none"> 사진, 영상, 음성 등 미디어 정보 	<ul style="list-style-type: none"> 사진, 위성사진, 항공사진, 영상, 음성 등
센서 데이터 (Sensor data)	<ul style="list-style-type: none"> 추적장치정보(Tracking device data)를 포함한 각종 센서를 통하여 생성되는 정보 	<ul style="list-style-type: none"> 위치, 기상, 수질, 대기, IoT 등
활동 데이터 (Online behavior data)	<ul style="list-style-type: none"> 의견정보(Opinion data), 웹 검색정보 등을 포함한 온라인 상에서 생성된 것으로 분석을 위하여 가공된 정보 	<ul style="list-style-type: none"> 인터넷 검색 및 페이지뷰 인덱싱 정보(WebLog) 등 이메일, 카카오톡, 트위터 등 공개 자료

(*진단 대상 분류는 플랫폼 자체적으로 마련 가능하다. 단, 가이드에서 제시하는 분류와 1:1 또는 1:N(상세분류)으로 매핑 되어야 한다.)

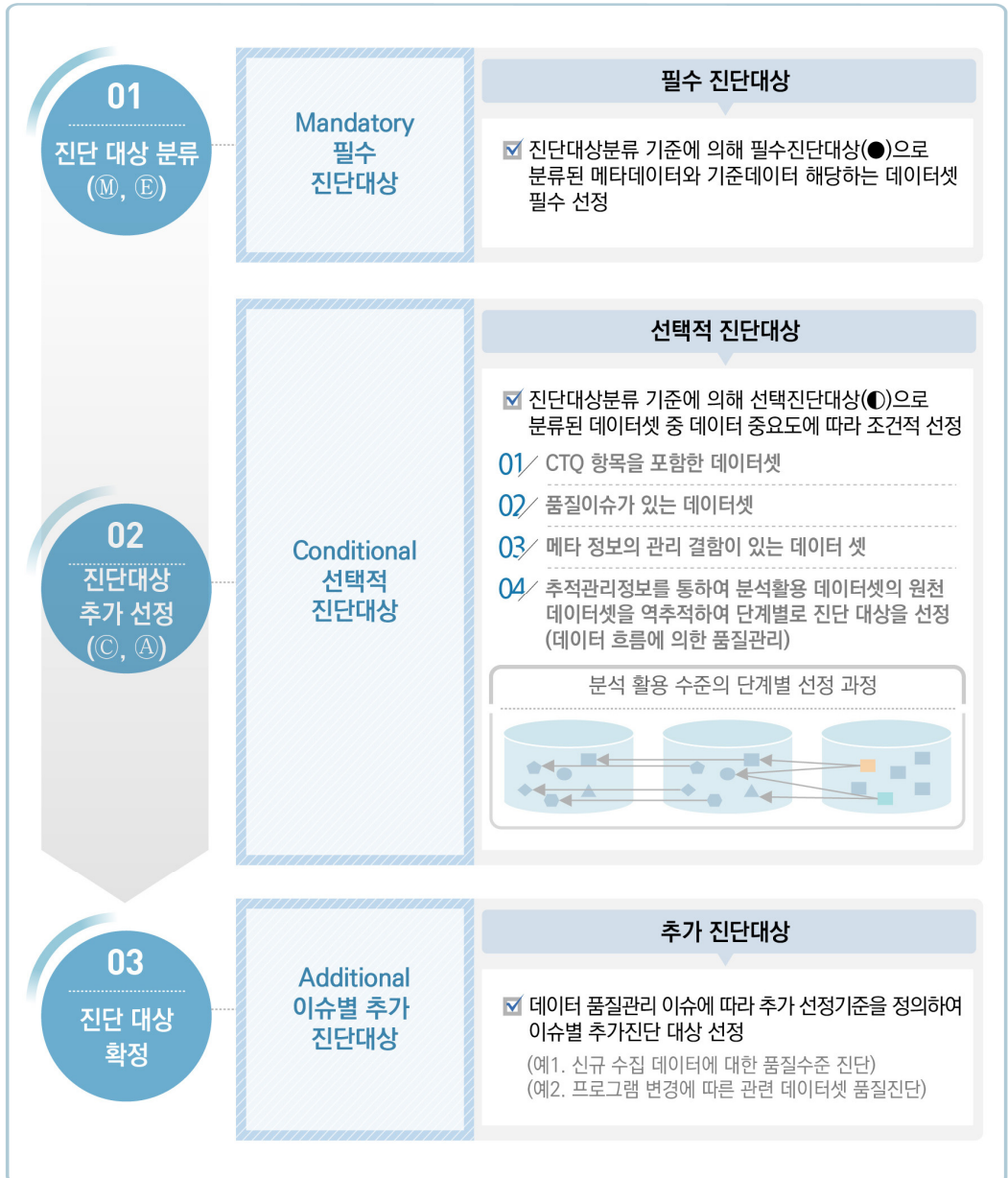
표 17 진단 대상 유형분류 작성 예시

주제영역	테이블명	테이블한글명	유형분류
회원	T_CUSTOMER_M	고객기본	기준
SNS	T_SN_EMPLYM__WEB_DOC	고용웹문서정보	분석
기업	T_ENTRPRS_STTUS_STATS	기업현황통계	분석
숙박	T_DD_ACDT_RT_L	일일숙박대여내역	거래
지자체카드	T_LC_POLICY_WIK_SETLE_INFO	정책주간결제정보	거래

▶ 진단 대상 선정

진단대상 분류 기준에 의해 필수진단대상을 선정하고, 선택 진단대상 중 선정기준에(품질이슈, 메타정보품질수준, 분석활용수준 등)따라 평가 후 추가 선정하여 품질진단 대상을 확정한다.

그림 24 진단 대상 및 선정 기준



I

II

III

IV

데이터 품질진단 방법

V

표 18 진단 대상 선정 기준 설명

구분		정의	기준
Mandatory 필수 진단대상	메타데이터 기준데이터	<ul style="list-style-type: none"> 메타데이터와 기준데이터에 해당하는 데이터 셋 선정 	<ul style="list-style-type: none"> 데이터 유형분류에 의한 메타데이터와 기준데이터
Conditional 선택적 진단대상	CTQ 대상	<ul style="list-style-type: none"> 정의되어 관리되고 있는 CTQ에 해당하는 항목(컬럼)을 포함하고 있는 데이터셋 	<ul style="list-style-type: none"> CTQ 항목을 포함한 데이터셋
	분석활용 대상	<ul style="list-style-type: none"> 데이터의 활용도를 산정하여 활용도가 높은 데이터셋 <ul style="list-style-type: none"> - 분석·활용에 이용된 원천 데이터셋을 단계별로 역추적한 데이터셋의 이용 수 - 무/유상 제공 데이터셋의 유통횟수 	<ul style="list-style-type: none"> 유통(분석인용/유통)된 데이터셋
	품질이슈 대상	<ul style="list-style-type: none"> 품질이슈가 있는 데이터셋 	<ul style="list-style-type: none"> 품질이슈가 있는 데이터셋
	메타정보 누락, 오류 대상	<ul style="list-style-type: none"> 데이터셋의 메타정보에 대하여 품질수준을 측정하여 품질 이슈가 있는 데이터셋 <ul style="list-style-type: none"> - 데이터셋 표준화 메타 정보 누락, 오류 - 데이터셋에 대한 설명 메타 누락, 오류 	<ul style="list-style-type: none"> 메타정보 누락, 오류가 있는 모든 데이터셋 <ul style="list-style-type: none"> - 표준화 메타 누락, 오류 - 설명메타 누락, 오류
Additional 이슈별 추가 진단대상	이슈별 추가진단 대상	<ul style="list-style-type: none"> 데이터 품질관리 이슈에 따라 정의한 선정을 충족하는 데이터셋 <ul style="list-style-type: none"> - 예1. 신규 수집 데이터의 품질수준 진단 - 예2. 프로그램 변경으로 영향을 받는 데이터셋 품질진단 	<ul style="list-style-type: none"> 이슈에 따라 선정기준 정의 <ul style="list-style-type: none"> - 데이터 품질이슈에 따라 정의한 추가 선정기준을 충족하는 데이터셋

표 19 진단 대상 선정 작성 예시

주제영역	테이블명	테이블한글명	유형분류	진단대상선정여부
회원	T_CUSTOMER_M	고객기본	기준	Y
SNS	T_SN_EMPLYM__WEB_DOC	고용웹문서정보	분석	Y
기업	T_ENTRPRS_STTUS_STATS	기업현황통계	분석	Y
숙박	T_DD_ACDT_RT_L	일일숙박대여내역	거래	Y
지자체카드	T_LC_POLICY_WIK_SETLE_INFO	정책주간결제정보	거래	Y

▶ 대량 데이터 선정

대량 데이터의 효율적 품질진단을 수행하기 위하여 표본조사를 할 수 있다.

신뢰수준 99%, 표본오차 0.1%일 때 표본크기 167만은 모집단 크기와 관계없이 진단결과의 대표

성을 가지므로 표본의 크기는 167만 정도면 적절하다.

그림 25 대량 데이터 규모 크기 및 표본 크기

데이터 규모에 따른 표본크기의 비교							품질진단 시 표본크기	
구분	표본설계 기준						신뢰수준	99%
	신뢰수준 99%, 표본오차 0.1%						표본오차	0.1%
데이터 규모	무한대	100억	50억	10억	1억	3천만	데이터규모	제한없음
표본크기	166.4만	166.3만	166.3만	166.1만	163.6만	157.6만	표본크기	167만

표본 진단 활용 가이드

- 대량 데이터의 크기 기준을 3천만으로 제시함
- 실제 데이터의 규모가 일정하지 않아 데이터 규모에 따라 과소표본 추출의 방지 및 데이터 규모에 따른 표본 크기의 계산을 간소화하기 위하여 표본조사를 통한 품질진단 시 표본의 크기를 모집단이 무한대일 때 표본 크기(166.4만) 보다 큰 167만을 표본조사를 통한 품질진단 시 표본의 크기로 제시함

표 20 표본진단 관련 용어

용어	의미	예시
신뢰수준	<ul style="list-style-type: none"> • 신뢰 수준 99%는 해당 표본조사를 99% 믿을 수 있다는 뜻이 아니라 같은 조사를 100번 하면 오차범위 내 동일한 결과가 나올 가능성이 99번이라는 의미이다. 	<ul style="list-style-type: none"> • 제품의 품질조사에서 99% 신뢰수준이라고 함은, 100번 조사를 하였을 때 99번은 같은 결과가 나온다는 것으로, 이는 표본을 골고루 잘 선택하였다는 뜻이다.
표본오차	<ul style="list-style-type: none"> • 표본으로부터 계산되어지는 추정치와 모집단의 값과 차를 표본오차라 한다. 	<ul style="list-style-type: none"> • 품질검사를 실시하여 불량품이 없을 때를 0, 전부가 불량품일 때를 100으로 했을 때, 지금 조사한 불량률이 3%, 표본오차 $\pm 0.1\%$라고 한다면 불량률은 2.9%~3.1%의 범위에 존재한다.

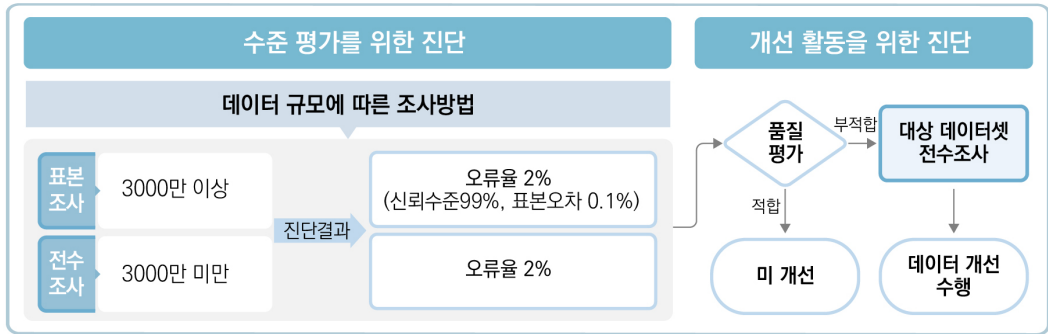
표 21 대량데이터 표본진단 작성 예시

주제영역	테이블명	테이블한글명	유형 분류	진단대상 선정여부	전수 /샘플
회원	T_CUSTOMER_M	고객기본	기준	Y	전수
SNS	T_SN_EMPLYM__WEB_DOC	고용웹문서정보	분석	Y	샘플
기업	T_ENTRPRS_STTUS_STATS	기업현황통계	분석	Y	전수
숙박	T_DD_ACDT_RT_L	일일숙박대여내역	거래	Y	전수
지자체카드	T_LC_POLICY_WIK_SETLE_INFO	정책주간결제정보	거래	Y	전수

▶ 대량 데이터의 품질진단 방안

대량 데이터의 기준 규모는 3천만 이상이면 표본을 167만으로 추출하여 표본조사를 수행할 수 있으며 표본조사 방법의 적용 여부는 자율적으로 결정하여 수행한다.

그림 26 대량 데이터 선정 절차



표본 진단 활용 가이드

- 표본 진단은 대량의 데이터를 관리하는 Big Table의 경우 품진 진단의 비용(시간, 자원)을 고려하여 효율적인 품질진단을 위하여 제시함
- 품질 수준의 평가를 위한 진단 수행 시 적용하는 것을 권장하며 개선을 위한 진단 시에는 반드시 전수조사를 시행하여야 함

04

항목(컬럼) 단위 선정



항목(컬럼) 단위 대상 선정은 기 선정 된 품질진단 대상 데이터셋(테이블)에서 품질진단 대상 도메인에 해당하는 항목을 필수 진단 대상으로 선정한다.

▶ 품질진단 대상 도메인

표 22 품질진단 대상 도메인 및 항목선정 예시

도메인	항목(컬럼) 선정 예시
코드	• 부동산분류코드, 고객구분코드, 처리 진행상태 코드 등
여부	• 사용여부, 존재유무 등
번호	• 고객번호, 우편번호, 사업자등록번호, 이메일주소 등 (형식 검증이 필요한 항목)
날짜	• 날짜(YYYYMMDD), 년월(YYYYMM), 년(YYYY), 월(MM), 일(DD) 일시(YYYYMMDDHH24MISS) 등
금액	• 매출금액, 급여액, 지급액, 공시지가 등
수량	• 가구 수, 인원수, 매도건수, 강수량, 몸무게, 온도, 압력 등
율	• 계약률, 이자율, 직군비율 등

※ 위의 7개 도메인에 해당하는 모든 컬럼은 반드시 진단 대상으로 선정하여 품질진단을 수행하여야 함

표 23 항목(컬럼)선정 작성 예시

테이블명	테이블한글명	컬럼명	도메인	진단대상선정여부
T_CUSTOMER_M	고객기본	고객명	명	N
T_CUSTOMER_M	고객기본	고객구분코드	코드	Y
T_CUSTOMER_M	고객기본	결혼여부	여부	Y
T_CUSTOMER_M	고객기본	결혼일자	일자	Y
T_DD_ACDT_RT_L	일일숙박대여내역	일일예약률	율	Y

I

II

III

IV

데이터 품질진단 방법

V

01

개요



데이터 품질진단 수행의 목적은 시스템에서 발생 되는 잠정 오류를 도출 및 확정하여 데이터 품질 수준을 파악하고자 하기 위함이다.

그림 27 품질진단 수행(Measure) 절차



표 24 품질진단 수행 절차

절차	내용
M1.품질진단 수행 계획	<ul style="list-style-type: none">품질진단 목적, 품질진단 범위 및 일정, 품질진단 환경구성 등을 고려하여 품질진단 계획을 수립한다.- 산출물 : 품질진단 수행계획서
M2.품질진단 규칙 정의	<ul style="list-style-type: none">프로파일 분석, 메타표준정보 분석, 업무규칙 등으로 품질진단을 하기 위한 진단규칙을 정의한다.- 산출물 : 프로파일링 결과내역, 품질진단 규칙 정의내역
M3.품질진단 실시	<ul style="list-style-type: none">정의 된 품질진단 규칙으로 품질진단을 실시하며, 진단결과 검토 과정을 통해 오류데이터 확정 및 품질진단 규칙 재정의 여부를 판단한다. 이 과정은 최종 오류데이터 확정시까지 “품질진단규칙정의”품질진단 실시/검토 과정을 반복 수행한다.
M4.품질진단 결과	<ul style="list-style-type: none">최종 오류 확정 내역(데이터셋 수, 항목 수, 전체건수, 오류건수, 오류율 등)을 값, 구조, 표준, 관리체계 관점에서 품질진단 결과서를 작성한다.- 산출물 : 품질진단 결과내역, 품질진단 결과서

02

품질진단 수행 계획



품질진단 수행계획은 진단 규칙 정의 및 실시에 앞서 계획을 수립함으로써 효과적인 품질진단의 목적을 달성 할 수 있도록 진단의 목적 및 목표, 일정, 환경구성 등을 정의하는 준비 절차이다.

품질관리 담당자는 품질진단수행계획서를 작성하여 품질진단 수행 전 관련 이해관계자들에게 공유한다.

표 25 품질진단 수행 계획서 구성

항 목	내 용
품질진단 목적	<ul style="list-style-type: none"> 품질진단을 수행하는 목적에 대하여 기술한다. 예) 품질진단 정책에 따른 월 정기 1차 데이터 품질진단을 수행
품질진단 대상	<ul style="list-style-type: none"> 품질진단 대상의 시스템 정보, 진단 대상, 제외 대상에 대하여 기술한다. - 저장영역, DBMS, 주제영역, 전체 테이블 컬럼, 진단 대상, 진단 제외 대상 정보 항목 등
품질진단 환경구성	<ul style="list-style-type: none"> 품질진단을 실시하는 환경에 대해 기술한다. 예) 운영데이터 품질진단 전용 저장소로 옮겨서 실시, 품질진단 방식(SQL 작성, DQ TOOL)의 설명 등
품질진단 일정	<ul style="list-style-type: none"> 품질진단 수행 단계 별 일정을 상세히 기술한다. - 품질진단환경구성, 품질진단규칙정의, 품질진단수행, 품질진단결과 확인 항목의 수행일정 등

I

II

III

IV

데이터 품질진단 100제

V

03

품질진단 규칙 정의



프로파일 진단기법과 업무규칙 진단기법 등을 사용하여 데이터 품질을 진단하기 위한 진단규칙을 정의한다.

▶ 프로파일 진단기법

- 데이터 현황을 파악하기 위해 통계적 분석(NULL값 수 · 공백값 수 · MIN값 · MAX값 등 산출)을 통해 잠재적 오류 징후를 발견하는 기법
- 메타정보(산출물 등)를 분석하여 잠재적 오류 징후를 발견하는 기법
- 비즈니스적 측면이 아닌 기술적 분석 측면을 기준으로 측정

▶ 업무규칙 기법

- 프로그램 분석 또는 업무담당자의 업무설명을 통해 업무규칙 도출하여 진단하는 기법
- 기술적 분석(프로파일링) 관점이 아니고 업무적 관점(업무 처리 규정 및 절차)으로 진단하는 기법

▶ 진단규칙 정의 시 유의사항

진단 스크립트(SQL)는 오류율 계산에 필요한 총 건수/오류건수를 구하는 스크립트와 진단규칙에 해당하는 오류데이터를 검색하는 스크립트 등 3종을 작성한다.

진단 스크립트(SQL)는 오류데이터를 검색 할 수 있도록 부정형으로 작성한다.

3.1 프로파일 진단규칙

진단대상 항목의 유효값, 유효범위를 메타정보 분석 또는 통계적 분석(총건수, null값, 최소값, 최대값 등 산출)으로 오류로 예상되는 항목에 대해 진단규칙을 도출한다.

표 26 프로파일 진단규칙 도출 절차

절 차	내 용	비 고
분석	메타정보 수집/분석	<ul style="list-style-type: none"> 메타정보(시스템 카탈로그)와 데이터 관리 문서 수집하여 누락, 불일치 요소들을 확인 정의 된 메타표준정보로 진단규칙 도출이 불가 할 경우 “프로파일 분석”이 필요함
	항목(컬럼) 분석	<ul style="list-style-type: none"> 항목(컬럼)속성분석을 통해 유효값, 유효범위 등을 판단 (MIN / MAX값, NULL / SPACE 분석, 최소빈도/ 최대빈도 패턴 등) 정의 된 표준정보 기반으로 진단규칙 도출
도출	진단규칙 도출	<ul style="list-style-type: none"> “프로파일 분석”결과를 통한 유형 별 품질진단규칙 도출 정의 된 표준정보 기반으로 진단규칙 도출 프로파일 분석 결과로부터 진단 규칙 도출

3.1.1 메타 · 표준정보 수집/분석

데이터관리 관련 산출물과 시스템 카탈로그(물리 정보)의 비교를 통해 품질진단대상의 현행화 수준 및 비표준을 파악하기 위해 실시한다. 비표준 사항 파악은 구조, 표준 진단의 참고사항으로 활용 할 수 있다. 데이터관리 관련 산출물로는 ‘데이터셋(테이블) 정의서’, ‘항목(컬럼) 정의서’, ‘도메인 정의서’, ‘ERD 관계(RI) 정의서’ 등이 있다.

아래 그림과 같이 데이터 관리 문서와 시스템 카탈로그(물리 정보) 비교를 통해 항목(컬럼)명, 데이터 타입, 길이 등이 불일치한 것을 발견할 수 있다.

그림 28 비표준사항 발견 예시

DBMS 메타 정보 (예시)

테이블영문명	테이블한글명	컬럼순서	컬럼영문명	컬럼한글명	데이터타입	길이	넓이부	PK
CUSTOMER	고객	1	CUST ID	고객번호	VARCHAR2	5	N	PK
CUSTOMER	고객	2	CUST DVCD	고객구분코드	VARCHAR2	5	Y	
CUSTOMER	고객	3	JUMIN RGST NO	고객주민등록번호	VARCHAR2	15	Y	
CUSTOMER	고객	4	BS RGST NO	사업자번호	VARCHAR2	15	Y	
CUSTOMER	고객	5	CUST STCD	고객상태코드	VARCHAR2	5	Y	
CUSTOMER	고객	6	CUST RGST DT	고객등록일자	VARCHAR2	8	Y	
CUSTOMER	고객	7	CUST NM	고객명	VARCHAR2	10	Y	
CUSTOMER	고객	8	COMB RES YN	통합고객등록여부	VARCHAR2	1	Y	

비교

데이터셋(테이블)/항목(컬럼)정의서 (예시)

테이블명	컬럼영문명	컬럼한글명	데이터타입	길이	넓이부	PK
고객 CUSTOMER	CUST ID	고객번호	VARCHAR2	5	N	PK
	CUST DVCD	고객구분코드	VARCHAR2	5	Y	
	JUMIN RGST NO	고객주민등록번호	NUMBER	13	Y	
	BS RGST NO	사업자번호	VARCHAR2	15	Y	
	CUST STCD	고객상태코드	VARCHAR2	5	Y	
	CUST NAME	고객명	VARCHAR2	10	Y	
	COMB RES YN	통합고객등록여부	VARCHAR2	1	Y	

불일치 사항 발견 (예시)

테이블영문명	컬럼영문명	데이터타입	길이	넓이부	PK	일치 여부	불일치사항	비고
고객 CUSTOMER	CUST ID	VARCHAR2	5	N	PK	일치		
	CUST DVCD	VARCHAR2	5	Y		일치		
	JUMIN RGST NO	VARCHAR2	15	Y		불일치	데이터타입 NUMBER(13)	
	BS RGST NO	VARCHAR2	15	Y		일치		
	CUST STCD	VARCHAR2	5	Y		일치		
	CUST RGST DT	VARCHAR2	8	Y		불일치	컬럼누락	
	CUST NM	VARCHAR2	10	Y		불일치	컬럼명 CUST NAME	
	COMB RES YN	VARCHAR2	1	Y		일치		

3.1.2 항목(컬럼) 분석

통계 함수(총건수, Null, 최대값, 최소값 등)를 사용하여 오류로 예상되는 항목을 분석하는 단계로 진단규칙유형(여부, 범위, 형식, 날짜 등)을 도출하기 위한 준비 과정 이다. 도메인별 항목(컬럼) 분석 방법 및 예시는 아래 표 및 예시를 참고하도록 한다.

표 27 항목(컬럼) 도메인별 분석 방법

도메인	항목(컬럼) 분석 방법	분석 유형(예)
코드	• 코드 도메인 항목에 대해 DISTINCT분석을 통하여 코드 값의 생성 분포를 파악	DISTINCT
여부	• 여부 도메인 항목에 대해 DISTINCT분석을 통하여 코드 값의 생성 분포를 파악	DISTINCT
번호	• 번호 등 채번 규칙이 있는 항목에 대해 채번 형식을 파악	PATTERN
날짜	• 날짜 도메인 항목에 대해 정의된 날짜 형식을 파악	PATTERN
금액	• 금액 도메인 항목에 대해 MIN, MAX값 파악	MIN, MAX
수량	• 수량 도메인 항목에 대해 MIN, MAX값 파악	MIN, MAX
율	• 율(비율) 도메인 항목에 대해 MIN, MAX값 파악	MIN, MAX

▶ 항목(컬럼) 분석(예시)

▶ NULL, SPACE 분석

- NULL 값과 SPACE 건수를 통하여 값의 유무를 파악 할 수 있다.
- 컬럼의 사용유무, 필수 입력 값 등에 대한 오류 징후를 파악 할 수 있다.

테이블명	컬럼명	TOT	NULL	SPACE
공급업체	기타정보	240	220	20

(분석 SQL - Oracle 11g 기준)

```
SELECT COUNT(*)          AS TOT_CNT
      , COUNT(기타정보) AS NULL_CNT
      , COUNT(CASE WHEN 기타정보 IS NOT NULL
                  AND TRIM(기타정보) IS NULL
                  THEN '1'
                END)      AS SPACE_CNT
FROM 공급업체;
```

▶ MIN, MAX 분석

- MIN 값과 MAX 값을 통하여 수치 값의 범위 파악
- 개별 빈도 값을 추출하지 않고 빠른 시간 내에 오류징후 파악 가능

테이블명	컬럼명	MIN	MAX
주문 상세	판매금액	-9999999	999999999

(분석 SQL - Oracle 11g 기준)

```
SELECT '판매금액'
```

```

, MIN(판매금액) AS MIN
, MAX(판매금액) AS MAX
FROM 주문상세;

```

▶ DISTINCT 분석

- 중복값의 건수를 통해 값의 분포 파악
- 여부, 코드 등 값의 허용범위 분석으로 오류 징후 파악 가능

테이블명	컬럼명	DISTINCT	건수
고객	통합고객여부	Y	250
		N	245
		0	1
		1	20

(분석 SQL - Oracle 11g 기준)

```

SELECT   통합고객여부 AS DISTINCT
        , COUNT(*)    AS DISTINCT_CNT
FROM     고객
GROUP BY 통합고객여부;

```

▶ 패턴 분석

- 문자열 패턴 값의 분포를 파악하여 오류징후 파악 가능

테이블명	컬럼명	패턴 값	패턴 건수
주문	주문번호	A999999999	3500
		A999999	4

(분석 SQL - Oracle 11g 기준)

```

SELECT TRANSLATE (주문번호,
                  '0123456789ABCDEFGHIJKLMNOPQRSTUVWXYZ',
                  '9999999999AAAAAAAAAAAAAAAAAAAAAAAAAAAA' ) as PATTERN
, COUNT (*) AS CNT
FROM   주문
GROUP BY TRANSLATE (주문번호,
                    '0123456789ABCDEFGHIJKLMNOPQRSTUVWXYZ',
                    '9999999999AAAAAAAAAAAAAAAAAAAAAAAAAAAA' );

```

3.1.3 프로파일 진단규칙 도출

메타정보 분석 또는 통계적 항목분석을 통해 값/표준/구조에 대해 잠재적 오류 징후를 발견하기 위해 진단규칙을 도출한다. 품질지표 별 진단방법 및 진단 규칙 도출 예시는 아래 표 및 예시를 참고하도록 한다.

표 28 품질지표 별 진단방법 및 진단 규칙 도출 방법(프로파일)

품질지표		진단내용	진단방법 및 진단규칙 도출 예시
완전성	구조 완전성	테이블 중복 진단	<ul style="list-style-type: none"> 전체 테이블 중에서 테이블의 컬럼명, 데이터 타입, 데이터 길이가 100% 동일한 테이블이 있는지 진단
		테이블 현행화 진단	<ul style="list-style-type: none"> 데이터모델에 정의되지 않았으나 실제DB에만 존재하는 경우, 데이터모델에 정의되었으나 실제DB에는 존재하지 않는 경우의 불일치 테이블 진단(테이블명)
		컬럼 현행화 진단	<ul style="list-style-type: none"> 데이터모델에 정의되지 않았으나 실제DB에만 존재하는 경우, 데이터모델에 정의되었으나 실제DB에는 존재하지 않는 경우의 불일치 컬럼 진단(컬럼명, Type, 길이, PK, Not Null)
		식별자 누락 진단	<ul style="list-style-type: none"> 전체 테이블 중에서 식별자(PK)가 정의되어 있지 않은 테이블이 있는지 진단
유효성	코드 유효성	코드 유효값 진단	<ul style="list-style-type: none"> 코드 도메인 항목의 값이 표준코드에 정의된 유효 코드 값과 일치하는지 진단 - 부동산분류코드는 공통코드에서 관리되는 값의 범위가 아니면 오류
	형식 유효성	패턴 진단	<ul style="list-style-type: none"> 정해진 형식(패턴)을 준수하는지 진단 - 고객번호 항목의 값은 6자리 숫자로 구성 되어 있지 않으면 오류
	여부 유효성	여부 유효값 진단	<ul style="list-style-type: none"> 여부 도메인 항목의 값이 정해진 값으로 관리되고 있는지 진단 - 여부 항목의 값이 'Y','N'이 아니면 오류)
	날짜 유효성	날짜형식 진단	<ul style="list-style-type: none"> 일자 도메인 항목의 값이 해당 도메인의 일자 형식을 준수하는지 진단 - 시작일자는 'YYYYMMDD' 이외 형식의 값은 오류
	범위 유효성	금액	<ul style="list-style-type: none"> 부동산 공시지가는 평당 1,000~100,000,000 사이의 값이 아니면 오류
		율	<ul style="list-style-type: none"> 부동산 계약률은 0~100% 사이의 값이 아니면 오류
		수량	<ul style="list-style-type: none"> 미분양 가구수는 0~10,000 사이의 값이 아니면 오류
일관성	표준 일관성	표준용어 진단	<ul style="list-style-type: none"> 모델(구조)의 모든 컬럼을 대상으로 표준용어를 준수하지 않은 컬럼(명)이 있는지 진단
		표준도메인 진단	<ul style="list-style-type: none"> 모델(구조)의 모든 컬럼을 대상으로 표준도메인을 준수하지 않은 컬럼(type,길이)이 있는지 진단
		엔티티 명명규칙 진단	<ul style="list-style-type: none"> 모델(구조)의 엔티티 명이 명명규칙을 준수하지 않은 엔티티(한글명)가 있는지 진단
		테이블 명명규칙 진단	<ul style="list-style-type: none"> 모델(구조)의 테이블명이 명명규칙을 준수하지 않은 테이블(명)이 있는지 진단
정확성	메타 정확성	비정형 속성정보의 정확성 진단	<ul style="list-style-type: none"> 비정형 콘텐츠의 메타 정보가 해당 비정형 콘텐츠의 속성 정보를 관리하는 컬럼의 데이터 값과 불일치하는지 진단

▶ 코드 유효성 진단규칙 도출 예시

1. 진단 대상 선정

- 주문 테이블 '주문상태코드' 항목 진단 규칙 도출 예시(공통코드)

2. 프로파일 분석

2.1 메타정보(공통코드 테이블)

테이블명	분류코드	분류코드명	코드값	코드값명
공통코드	1001	주문상태코드	01	주문취소
공통코드	1001	주문상태코드	02	주문완료

(분석 SQL - Oracle 11g 기준)

```
SELECT *
FROM   공통코드
WHERE  분류코드명 = '주문상태코드';
```

2.2 항목(컬럼) 분포 값 조회(DISTINCT)

테이블명	컬럼명	분포 값(DISTINCT)	분포 건수
주문	주문상태코드	01	50
		02	1200
		03	2

(분석 SQL - Oracle 11g 기준)

```
SELECT   주문상태코드
        , COUNT(*)
FROM     주문
GROUP BY 주문상태코드;
```

3. 진단대상 분석 결과 및 진단규칙 도출

- '주문상태코드' 항목의 값 분포 및 공통코드 테이블의 데이터 조회를 통해 유효값을 파악 할 수 있다.
- 항목(컬럼) 분포 값 조회 확인 결과 '01', '02' 값 이외의 데이터는 오류로 예측 할 수 있다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT *
FROM   주문 LEFT OUTER JOIN 공통코드
        ON ('1001' = 공통코드.분류코드
            AND 주문.주문상태코드 = 공통코드.코드값)
WHERE  공통코드.코드값 IS NULL
GROUP BY 주문.주문상태코드;
```

여부 유효성 진단규칙 도출 예시(프로파일링)

- 진단 대상 선정
 - 고객 테이블의 '결혼여부' 항목 진단 규칙 도출 예시

2. 프로파일 분석

2.1 메타정보(도메인 정의서)

도메인명	TYPE	길이	정의
여부	VARCHAR2	1	유효값은 'Y' , 'N'으로 정의한다.

2.2 항목(컬럼) 분포 값 조회(DISTINCT)

테이블명	컬럼명	분포 값(DISTINCT)	분포 건수
고객	결혼	Y	650
		N	400
		X	12

```

(분석 SQL - Oracle 11g 기준)
SELECT  결혼여부, COUNT(*) AS CNT
FROM    고객
GROUP BY 결혼여부;
    
```

- 진단대상 분석 결과 및 진단규칙 도출
 - 중복값의 건수를 통해 값의 분포를 파악 할 수 있다.
 - 확인 결과 'Y','N'값 이외의 데이터는 오류로 예측할 수 있다.

```

(진단규칙 SQL - Oracle 11g 기준)
SELECT  *
FROM    고객
WHERE   결혼여부 NOT IN ('Y','N');
    
```

▶ 형식 유효성 진단규칙 도출 예시(프로파일링)

1. 진단 대상 선정

- 장애기본 테이블의 '장애접수번호' 항목 진단 규칙 도출 예시

2. 프로파일 분석

2.1 메타정보(채번규칙)

채번규칙명	채번규칙
장애접수번호	C(영문 1자리) + 장애접수일자(YYYYMMDD) + '-' + 숫자 6자리

2.2 항목(컬럼) 분포 값 조회(PATTERN)

테이블명	컬럼명	패턴 값(PATTERN)	패턴 건수
장애기본	장애접수번호	A99999999-999999	2
		C99999999-999999	800

(분석 SQL - Oracle 11g 기준)

```
SELECT TRANSLATE (장애접수번호,
                  'C-0123456789', 'C-9999999999' ) as PATTERN
       ,COUNT (*) AS CNT
FROM   주문
GROUP BY TRANSLATE (장애접수번호,
                   'C-0123456789', 'C-9999999999' );
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 채번규칙과 패턴 값의 분포 건수를 통해 값의 분포를 파악 할 수 있다.
- 확인 결과 장애접수번호 채번규칙 이외의 데이터는 오류로 예측할 수 있다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT   장애접수번호
        , 장애접수일자
FROM     장애기본
WHERE    SUBSTR(장애접수번호,1,1) <> 'C'
OR       SUBSTR(장애접수번호,2,8) <> 장애접수일자
OR       SUBSTR(장애접수번호,10,1) <> '-' ;
```

I

II

III

IV

데이터 품질진단 방법론

V

▶ 날짜 유효성 진단규칙 도출 예시(프로파일링)

1. 진단 대상 선정

- 고객 테이블의 '가입일자' 항목 진단 규칙 도출 예시

2. 프로파일 분석

2.1 메타정보(도메인정의서)

도메인명	TYPE	길이	정의
일자	VARCHAR2	8	일자의 유효범위 값은 YYYYMMDD 이다.

2.2 항목(컬럼) 분포 값 조회(PATTERN)

테이블명	컬럼명	TYPE	길이	패턴 값(PATTERN)
고객	가입일자	VARCHAR2	20	99999999
				9999년9월9일

(분석 SQL - Oracle 11g 기준)

```
SELECT TRANSLATE (가입일자,
                  '0123456789', '9999999999' ) as PATTERN
, COUNT (*) AS CNT
FROM   고객
GROUP BY TRANSLATE (가입일자,
                  '0123456789', '9999999999' );
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 가입일자 항목의 패턴 값의 분포를 파악하여 일자 도메인 형식을 준수하지 않는 데이터는 오류로 예측 할 수 있다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT   고객번호,   가입일자
FROM     고객
WHERE    chkDate(가입일자) != 'Y';
```

※ chkData : input이 날짜 값(YYYYMMDD)으로 유효하면 'Y' 그렇지 않으면 'N'을 반환하는 사용자 함수이다.

▶ 범위 유효성 진단규칙 도출 예시(프로파일링)

1. 진단 대상 선정

- 고객 테이블의 '체중' 항목 진단 규칙 도출 예시

2. 프로파일 분석

2.1 메타정보(컬럼정의서)

테이블명	컬럼명	...	설명
고객	체중	...	체중항목의 유효범위는 1 ~ 200 이다.

2.2 항목(컬럼) 분포 값 조회(MIN/MAX)

테이블명	컬럼명	TYPE	MIN(체중)	MAX(체중)
고객	체중	NUMBER	-64.5	120

(분석 SQL - Oracle 11g 기준)

```
SELECT MIN(체중)
      , MAX(체중)
FROM 고객;
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 체중 항목의 MIN/MAX 값의 범위를 파악하여 체중 항목의 유효 범위에 해당하지 않는 데이터는 오류로 예측 할 수 있다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  고객번호, 체중
FROM    고객
WHERE   NOT 체중 BETWEEN 1 AND 200;
```

I

II

III

IV

데이터 품질진단 방법론

V

3.2 업무 진단규칙

▶ 진단규칙 도출

업무 관점의 품질이슈나 업무 담당자 인터뷰를 통해 값에 대한 진단 요건을 확보하여 진단 규칙을 도출한다. 품질지표 별 진단방법 및 진단 규칙 도출 예시는 아래 표 및 예시를 참고하도록 한다.

표 29 진단방법 및 진단규칙 도출 예시(업무규칙)

품질지표		진단내용	진단방법 및 진단규칙 도출 예시
완전성	단독 완전성	NULL값 진단	<ul style="list-style-type: none"> 업무상 필수로 입력되어야 하는 컬럼의 값이 Null인지 진단 - 고객기본 테이블의 고객명 항목의 값은 Not Null 이어야 한다.
	조건 완전성	NULL값 진단	<ul style="list-style-type: none"> 단일/복수 테이블의 컬럼 간 조건에 따른 필수 값이 Null 인지 진단 - 고객기본 테이블의 탈퇴일자 항목에 값이 있으면 탈퇴사유 항목에도 값이 있어야 한다.
유효성	선후 관계 유효성	업무적 선후관계 유효값 진단	<ul style="list-style-type: none"> 시간순서 또는 사건순서가 정해진 선후관계를 지켜서 값이 관리되고 있는지 진단(시간순서 유효성, 사건순서 유효성) - 시작일자는 종료일자 보다 빠른 일자 값이어야 한다.
일관성	참조 무결성	참조 무결성 진단	<ul style="list-style-type: none"> 참조관계가 있는 테이블의 참조항목(FK)의 값이 서로 존재해야 하며 일관 되게 관리되고 있는지 진단. (메타정보가 잘 정의되어 있으면 프로파일 진단 가능) - 교육신청내역 테이블의 고객번호는 고객기본의 고객번호 값에 존재해야 한다. - 데이터모델(ERD)의 참조 관계가 정의되었으나 실제DB에는 참조관계 무결성 제약이 정의되지 않은 경우 진단
	중복 일관성	중복 일관성 진단	<ul style="list-style-type: none"> 성능상의 이유로 중복하여 정의한 항목 중 원천 데이터와 불일치 한 값이 관리되고 있는지 진단. (메타정보가 잘 정의되어 있으면 프로파일 진단 가능) - 교육신청내역 테이블에서 관리되는 고객명은 고객기본 테이블에서 관리하는 고객명과 일치해야 한다.
정확성	논리 관계 정확성	업무규정에 따른 데이터 값 진단	<ul style="list-style-type: none"> 논리관계가 있는 항목에 대해 논리에 맞게 값이 관리 되고 있는 진단. - 고객구분이 '기업'이면 사업자등록번호의 값은 10자리 숫자 값으로 사업자번호 부여 규칙에 맞게 관리 되어야 한다.
	파생 항목 정확성	파생 데이터 값 정확성 진단	<ul style="list-style-type: none"> 파생되어 관리되는 항목에 대해 원천 항목의 값과 파생 된 항목의 값이 정확하게 관리되고 있는지 진단. (합계에 의한 파생, 새로운 항목 도출에 의한 파생) - 나이 항목의 값은 생년월일의 년도에 맞는 정확한 값으로 관리되어야 한다.

▶ 단독완전성 업무규칙 도출 예시

1. 진단 대상 선정

- 고객 테이블 '전화번호' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 고객 테이블 '전화번호' 항목은 고객정보 등록 시 반드시 입력되어야 한다.

2.2 진단 대상 데이터 조회

고객번호	성명	전화번호
E190115	박사원	
E180095	김사원	02-1234-5678
E200211	이사원	032-2345-6789

(조회 SQL - Oracle 11g 기준)

```
SELECT  고객번호, 성명
        , 전화번호
FROM    고객;
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 업무 처리 규정을 기반으로 진단 대상 데이터 조회 결과 '전화번호' 항목에 Null인 값이 존재함을 볼 수 있다. 따라서 Null인 데이터는 오류로 예측 할 수 있다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  고객번호
        , 성명
        , 전화번호
FROM    고객
WHERE   전화번호 IS NULL;
```

I

II

III

IV

데이터 품질진단 방법론

V

▶ 조건완전성 업무규칙 도출 예시

1. 진단 대상 선정

- 고객 테이블 '이메일주소' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 고객 테이블 '이메일주소' 항목은 고객정보의 마케팅수신여부가 'Y'일 경우에 반드시 입력되어야 한다.

2.2 진단 대상 데이터 조회

고객번호	마케팅수신여부	이메일주소
E190115	Y	
E180095	N	ABC123@GMAIL.COM

(조회 SQL - Oracle 11g 기준)

```
SELECT  고객번호, 마케팅수신여부
        , 이메일주소
FROM    고객;
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 업무 처리 규정을 기반으로 진단 대상 데이터 조회 결과 마케팅수신여부가 'Y'이나 이메일주소가 없는 데이터가 존재함을 볼 수 있다. 따라서 고객 테이블의 마케팅수신여부 항목의 값이 'Y'인 건 중 이메일주소의 값이 Null인 건은 오류로 예측한다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  고객번호, 마케팅수신여부, 이메일주소
FROM    고객
WHERE   마케팅수신여부='Y'
AND     이메일주소 IS NULL;
```

▶ 선후관계 유효성 업무규칙 도출 예시

1. 진단 대상 선정

- 고객 테이블 '가입일자', '해지일자' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 고객 테이블 '가입일자', '해지일자' 항목은 해지일자에 값이 있는 건 중 해지일자가 가입일자 보다 과거일 수 없다.

2.2 진단 대상 데이터 조회

고객번호	가입일자	해지일자
E190115	20190705	20190704
E200211	20191112	20191205

(조회 SQL - Oracle 11g 기준)

```
SELECT  고객번호
        , 가입일자
        , 해지일자
FROM    고객;
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 해지일자에 값이 존재하는 건 중 해지일자가 가입일자 보다 과거 값이 존재할 경우 오류로 예측한다.
(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  고객번호
        , 가입일자
        , 해지일자
FROM    고객
WHERE   해지일자 IS NOT NULL
AND     가입일자 > 해지일자 ;
```

I

II

III

IV
데이터 품질진단 방법론

V

▶ 참조 무결성 업무규칙 도출 예시

1. 진단 대상 선정

- 주문 테이블 '주문고객번호' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 주문 테이블 '주문고객번호' 항목의 값은 고객 테이블 '고객번호' 항목과 참조관계이어야 한다.
- 부모(고객번호) 없는 자식(주문고객번호)이 존재 할 수 없다.

2.2 진단 대상 데이터 조회

주문고객번호(주문 테이블)	고객번호(고객 테이블)
E190102	E190102
E190301	
E190613	E190613

(조회 SQL - Oracle 11g 기준)

```
SELECT  주문.주문고객번호
        , 고객.고객번호
FROM    주문 LEFT OUTER JOIN 고객
        ON  (주문.주문고객번호 = 고객.고객번호);
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 존재할 경우 오류로 예측한다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  주문.주문고객번호
        , 고객.고객번호
FROM    주문 LEFT OUTER JOIN 고객
        ON  (주문.주문고객번호 = 고객.고객번호)
WHERE   고객.고객번호 IS NULL;
```

▶ 중복 일관성 업무규칙 도출 예시

1. 진단 대상 선정

- 주문내역 테이블 '주문상태코드' 항목과 배송내역 테이블 '주문상태코드' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 주문내역 테이블 '주문상태코드' 항목과 배송내역 테이블 '주문상태코드' 항목의 값은 일관성이 유지되어야 한다.

2.2 진단 대상 데이터 조회

배송담당자	주문번호	주문상태코드 (주문내역 테이블)	주문상태코드 (배송내역 테이블)
이배송	P199-423	02	02
김배송	P203-111	02	01

(조회 SQL - Oracle 11g 기준)

```
SELECT  배송.배송담당자
        , 배송.주문번호
        , 배송.주문상태코드,
        주문.주문상태코드
FROM    배송내역 AS 배송 LEFT OUTER JOIN 주문내역 AS 주문
        ON (배송.주문번호 = 주문.주문번호)
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 주문, 배송 테이블의 주문상태코드 항목의 값이 불일치 할 경우 오류로 예측한다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  배송.배송담당자
        , 배송.주문번호
        , 배송.주문상태코드,
        주문.주문상태코드
FROM    배송내역 AS 배송 LEFT OUTER JOIN 주문내역 AS 주문
        ON (배송.주문번호 = 주문.주문번호)
WHERE   배송.주문상태코드 <> 주문.주문상태코드
```


▶ 논리관계 정확성 업무규칙 도출 예시

1. 진단 대상 선정

- 카드대금납부내역 테이블 '완납여부', '미납금액' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 카드대금납부내역 테이블의 완납여부가 'N'일 때 미납금액이 '0' 또는 완납여부가 'Y' 일때 미납금액이 '0'보다 큰 건이 존재 할 수 없다.

2.2 진단 대상 데이터 조회

고객번호	완납여부	미납금액
E190115	Y	0
E200211	N	0
E301111	Y	10,000

(조회 SQL - Oracle 11g 기준)

```
SELECT  고객번호, 완납여부, 미납금액
FROM    카드대금납부내역
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 카드대금납부내역 테이블의 완납여부가 'N'일 때 미납금액이 '0' 또는 완납여부가 'Y' 일때 미납금액이 '0'보다 큰 건이 존재하는 경우 오류로 예측한다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  고객번호, 완납여부, 미납금액
FROM    카드대금납부내역
WHERE   (완납여부 = 'N' AND 미납금액 = 0
        OR 완납여부='Y' AND 미납금액 <> 0 )
```

▶ 파생항목 정확성 업무규칙 도출 예시

1. 진단 대상 선정

- 청구 테이블 '청구금액', 카드사용내역 테이블의 '결제금액(월 집계)' 항목 진단 규칙 도출 예시

2. 업무규칙 분석

2.1 업무 처리 규정 및 관련 내용

- 청구 테이블 '청구금액' 값은 카드사용내역 테이블 '결제금액'의 월 집계 값과 일치해야 한다.

2.2 진단 대상 데이터 조회

고객번호	결제년월	결제금액합계	청구금액
E190115	201901	98,000	98,000
E190456	201902	77,000	77,000
E200211	201903	50,500	50,000
E301111	201908	32,000	33,000

(조회 SQL - Oracle 11g 기준)

```
SELECT 사용내역.고객번호, 사용내역.결제년월
    , 사용내역.금액합계, 청구.청구금액
FROM (
    SELECT 고객번호
        , SUBSTR(결제일자,1,6) AS 결제년월
        , SUM(결제금액)      AS 금액합계
    FROM 카드사용내역
    GROUP BY 고객번호, SUBSTR(결제일자,1,6)
    ) AS 사용내역
INNER JOIN 청구 AS 청구
ON (사용내역.고객번호 = 청구.고객번호
AND 사용내역.결제년월 = 청구.사용년월)
```

3. 진단대상 분석 결과 및 진단규칙 도출

- 청구 테이블 '청구금액' 값은 카드사용내역 테이블 '결제금액'의 월 집계 값과 일치하지 않는 경우 오류로 예측한다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT 사용내역.고객번호, 사용내역.결제년월
    , 사용내역.금액합계, 청구.청구금액
FROM (
    SELECT 고객번호
        , SUBSTR(결제일자,1,6) AS 결제년월
        , SUM(결제금액)      AS 금액합계
    FROM 카드사용내역
    GROUP BY 고객번호, SUBSTR(결제일자,1,6)
    ) AS 사용내역
INNER JOIN 청구 AS 청구
ON (사용내역.고객번호 = 청구.고객번호
AND 사용내역.결제년월 = 청구.사용년월)
WHERE 사용내역.금액합계 <> 청구.청구금액;
```

3.3 공간정보 진단규칙

3.3.1 공간정보의 구성 및 품질진단 대상

공간정보는 지리 정보 중에서 지표 공간 속에 분포하는 현상의 위치나 형태를 나타내는 정보를 공간 정보라고 한다. 공간정보는 점, 선, 면으로 표현된다. 상점이나 공장의 위치는 점으로, 도로와 하천, 교통·통신망의 이동은 선으로, 산업 단지나 주택 단지와 같이 활동의 영향권은 면으로 표현된다. 공간정보는 위도, 경도, 행정구역의 경계 등으로 표현할 수 있다.

공간정보는 지리적 위치를 갖고 있는 대상에 대한 공간 데이터(spatial data)와 속성자료(attribute data), 모델 구성 데이터, 공간정보 메타 데이터로 구성되어 있다.

그림 29 공간정보 데이터 구성 및 진단 대상

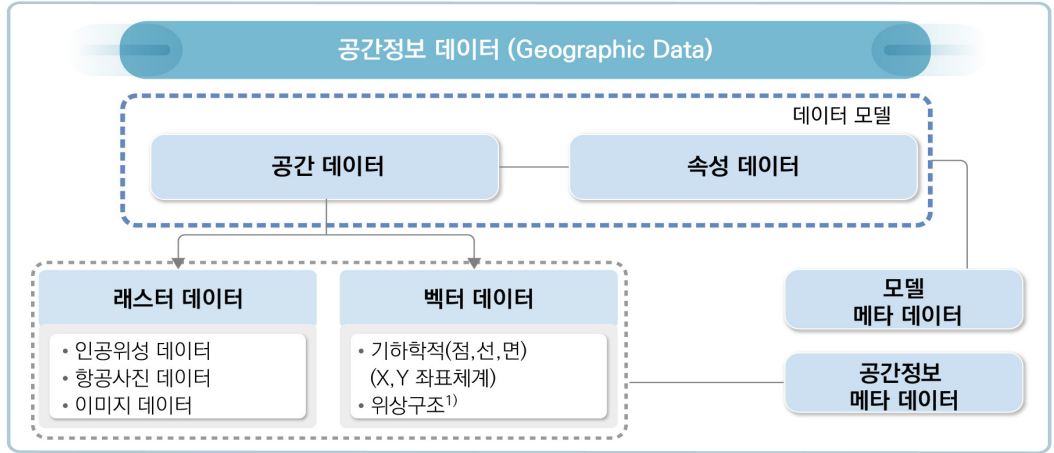


표 30 공간정보 구성요소의 설명

대상 구분	설명	예
공간 데이터 (Spatial data)	<ul style="list-style-type: none">• 지형지물의 공간특성을 표현한 데이터	<ul style="list-style-type: none">• 벡터 데이터, 래스터 데이터
속성 데이터	<ul style="list-style-type: none">• 어떤 대상의 주제나 비위치 특성을 나타내는 정보	<ul style="list-style-type: none">• 건물의 층수, 주소, 면적 등
모델 메타 데이터	<ul style="list-style-type: none">• 속성 데이터의 명명규칙, 표준사전, 표준도메인 등 표준화 관련 정보	<ul style="list-style-type: none">• 명명규칙, 표준사전, 표준도메인 등
래스터 데이터 (Raster data)	<ul style="list-style-type: none">• 셀마다 그 지리적 범위에 관한 특성값을 일정한 수치로 표시하는 격자 형식의 자료 구조를 가지는 자료	<ul style="list-style-type: none">• 인공위성 영상, CCD 사진기 영상, 또는 항공사진, 정사영상 등
벡터 데이터 (Vector data)	<ul style="list-style-type: none">• 모든 공간정보를 2차원 또는 3차원 좌표값으로 환산하여 점, 선, 면의 형태로 저장한 자료	<ul style="list-style-type: none">• Shape file 등
공간정보 메타 데이터	<ul style="list-style-type: none">• 공간정보를 설명하는 정보를 목록화하여 제공함으로써 사용자가 자료의 획득 및 사용에 도움을 주기 위하여 자료	<ul style="list-style-type: none">• 좌표계, 해상도, 축척, 도엽기준, 파일형식 등

3.3.2 진단규칙 도출

공간정보의 특성을 고려하여 구조화 기반 공간 진단, 레이어(layer) 기반 공간 진단, 육안검사 등을 통해 공간정보에 대해 잠재적 오류 징후를 발견하기 위한 진단규칙을 도출한다.

▶ 품질지표별 공간정보 진단방법

표 31 공간정보 품질진단 지표

품질지표		진단내용	진단방법
완전성	공간객체 완전성	공간객체 누락·초과	<ul style="list-style-type: none"> 두 레이어(layer) 비교하여 공간객체가 누락되어 있거나, 초과하는 객체들이 존재하는지 진단하는 지표
일관성	위상구조 일관성	공간정보 구조화	<ul style="list-style-type: none"> 공간객체가 기하학적 위상을 준수하여 생성되었는지를 진단
	개념적 일관성	개념적 일관성	<ul style="list-style-type: none"> 공간정보가 개념적 스키마 규칙(공간금칙규칙)에 위배된 공간객체가 존재하는지 진단
		개별공간파일 완결성	<ul style="list-style-type: none"> 개별 공간정보(레이어) 내의 공간객체들이 공간적 완결성(형식, 관계)을 유지하고 있는지에 대하여 진단. (동일한 공간유형, 중첩, 중복되지 않아야 함)
유효성	그리드 유효성	위치 정확성 확보	<ul style="list-style-type: none"> 래스터(Raster) 공간정보가 도엽 또는 격자 기준에 따른 래스터(Raster) 이미지의 픽셀수가 일치하는지를 진단(해상도 정확성 진단)
		범위 유효성	<ul style="list-style-type: none"> 도엽으로 이루어진 래스터(Raster) 공간정보의 공간적 범위가 도엽 범위와 일치하는지 여부를 진단
		공간참조체계 정확성	<ul style="list-style-type: none"> 좌표계(타원체/투영체계)가 정확한 값을 가지고 있으며, 공간정보가 해당 범위 내에 존재하는지를 진단
정확성	위치식별 정확성	식별자 정의	<ul style="list-style-type: none"> 기본공간정보를 기준으로 위치식별자(도로명주소, PNU(지번주소), UFID(공간정보참조체계) 등)가 정확하게 구축되어 있는지를 진단

I

II

III

IV

데이터 품질진단 방법

V

▶ 공간객체 누락·초과

표 32 공간객체 완전성

지표	완전성	세부지표	공간객체 완전성	진단규칙	공간객체 누락·초과
세부지표정의	• 공간객체가 누락되어 있거나, 초과하는 객체들이 존재하는지 진단				
진단 대상	• 품질진단 대상 공간정보 (점, 선, 면), 기준공간정보 ⁶⁾				• 점, 선, 면
진단 기법	• 레이어(layer)기반 공간진단				• 프로파일 진단
진단 조건	• 비교대상이 있는 공간정보에 국한하여 진단 ※ 토지, 건물, 도로, 하천 등 동일한 공간정보를 기관별로 각각 구축한 경우 진단				
진단 규칙	• 기준공간정보와 비교하여 누락된 공간 객체 및 추가로 생성된 공간객체가 없어야 한다.				
진단 설명	• 레이어 (layer) 기반 규칙 기능을 이용하여 연속지적도, 도로명주소, 국가기본도 등과 같은 기본공간정보를 기준으로 두 레이어 (layer)를 비교하여 존재하지 않으면 누락으로 진단하며, 기본공간정보에 없는 것이 존재하면 초과로 진단함 참고) ArcGis의 topology rule기능을 참조하여 진단 누락 : Must be larger than cluster tolerance 불일치 : Must cover each other				

▶ 공간객체 누락·초과 > 건물 일부 누락 예시

- 개별 공간정보에 대하여 동일위치의 공간객체들이 누락되어 있는 경우

그림 30 공간객체 완전성 예시(1)

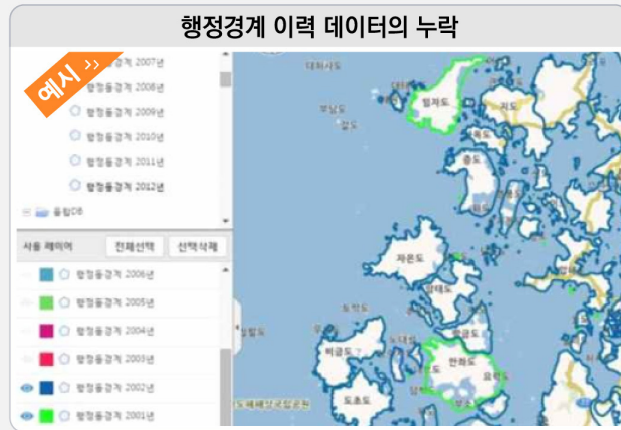


▶ 공간객체 누락·초과 > 행정경계의 일부 누락 예시

- 개별 공간정보에서 행정경계의 일부가 누락되는 경우

6) 기준공간정보: 기준이 되는 공간정보로 연속지적도, 도로명주소, 건물통합정보 등의 공신력 있는 공간정보

그림 31 공간객체 완전성 (예시 2)



▶ 위치식별 정확성

표 33 위치식별 정확성

지표	정확성	세부지표	위치식별 정확성	진단규칙	식별자 정의
세부지표정의	<ul style="list-style-type: none"> 기본공간정보를 기준으로 공간객체의 위치식별자가 일치하는지를 진단 				
진단 대상	<ul style="list-style-type: none"> 기준공간정보, 위치식별자 속성정보 				위치식별자 속성정보
진단 기법	<ul style="list-style-type: none"> 레이어(layer)기반 공간 진단 (2개 레이어 기반 진단) 				프로파일 진단
진단 조건	<ul style="list-style-type: none"> 위치식별자가 존재하는 공간정보 한하여 진단 				
진단 규칙	<ul style="list-style-type: none"> 기본공간정보를 기준으로 위치식별자(도로명주소, PNU(지번주소), UFID(공간정보참조체계) 등)가 정확하게 구축되어 있는지를 진단 예) PNU : 19자리(행정코드 10자리+식별코드 1자리+본번 4자리+부번 4자리) 예) UFID : 17자리(16진수) 				
진단 설명	<ul style="list-style-type: none"> 공간정보의 속성 항목의 위치식별자 정보들의 값이 기준공간정보의 해당 위치에 있는 공간객체가 관리하는 위치식별자 값과 일치하는지를 검사한다. 예) 생성한 공간객체의 속성값인 PNU(지번)주소가 연속지적도의 해당 공간위치의 PNU(지번)주소와 일치하는지를 검사한다. *) 해당 식별자의 정보들이 일반정보와 결합하여 활용할 수 있도록 하기 위함 				

I

II

III

IV

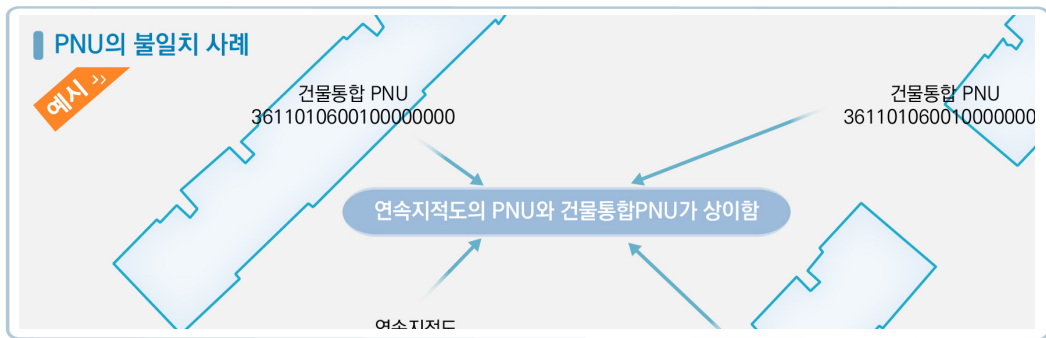
데이터 품질진단 방법

V

표 34 위치식별자 종류

구분	식별자	길이	형식	관련법
토지	지번 주소 (PNU)	19	<ul style="list-style-type: none"> • 법정동코드(10, 시도(2)시군구(3) 읍면동(3)리(2)) + 토지구분(1)*+ 지번본번(4) + 지번부번(4) • * 토지구분 : 1(일반), 2(산) 	<ul style="list-style-type: none"> • 공간정보의 구축 및 관리 등에 관한 법률 제 2 조, 동법 시행령 제 56 조
	도로명 주소	25	<ul style="list-style-type: none"> • 시군구코드(5)+도로코드(7) + 읍면동일련번호(2) + 지하여부(1)*+ 건물본번(5) + 건물부번(5) • * 지하여부 : 0(지상), 1(지하) 	<ul style="list-style-type: none"> • 도로명주소법 제 2 조 동법 시행령 제 3 조
	공간정보 참조체계 (UFID)	17	<ul style="list-style-type: none"> • 공간정보참조체계코드(분류체계(1)*+ 코드(7)) + 일련번호(8) + 오류점검수(1) • 교통(A), 건물(B), 시설(C), 식생(D), 수계(E), 지형(F), 경계(G), 기타(H) 	<ul style="list-style-type: none"> • 공간정보참조체계 부여·관리 등에 관한 규칙 제5조
건물	건물 식별번호	28	<ul style="list-style-type: none"> • *건물식별번호(건축물1동 단위) • *건축물대장 식별 유일번호(PK) 	<ul style="list-style-type: none"> • AIS 구축 이전 (1 억번 미만) /이후 (1 억번 부터 시작)
	건물통합정보 식별번호	28	<ul style="list-style-type: none"> • 건축물생성년도(4) + 위치정보 (x,y좌표값16) + 층별정보(5) + 호별정보 (3) 	-

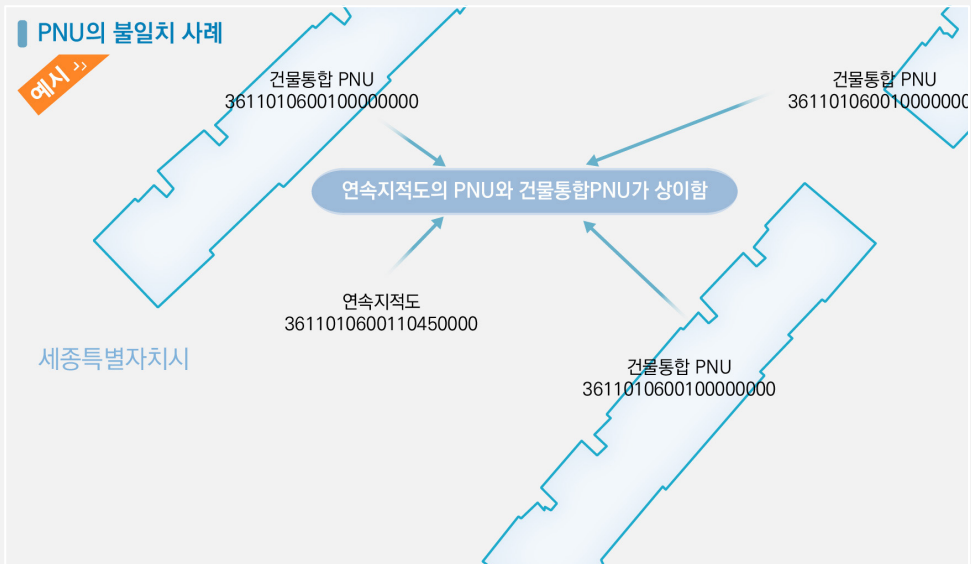
그림 32 위치 식별자



▶ 식별자 정의 > PNU의 불일치 예시

- 건물정보와 지적정보 내에 공통으로 존재하는 PNU가 일치하지 않는 경우가 발생할 수 있음

그림 33 위치식별 정확성 (예시)



I

II

III

IV
데이터 품질진단 방법론

V

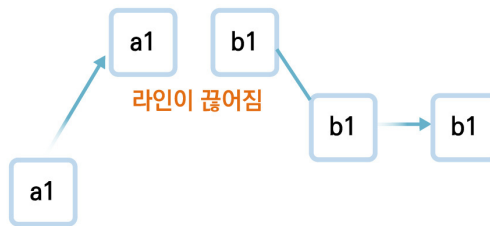
▶ 위상구조 일관성

표 35 위상구조 일관성

지표	일관성	세부지표	위상구조 일관성	진단규칙	공간정보 구조화
세부지표정의	공간객체가 기하학적 위상을 준수하여 생성되었는지를 진단하는 지표				
진단 대상	공간객체				점, 선, 면
진단 기법	구조화 기반 공간 진단				프로파일 진단
진단 조건	-				
진단 규칙	Check Geometry (기하학적 위상구조 11가지 오류 유형)				객체타입
	1101	Discontinuous parts	Part로 정의된 객체의 라인이 끊어지거나 불연속으로 구축된 오류. (Shape 파일로 최초 생성된 데이터의 진단 . DB로 구축된 경우는 진단제외)		선,면
	1102	Duplicate vertex	동일한 좌표에 두 개 이상의 중복 정점(Vertex)을 가지고 있는 오류		선,면
	1103	Empty parts	공간정보가 멀티 파트일 경우 그 중 하나가 비어 있는 오류 (Multi-Part 객체에 대한 진단)		점,선,면
	1104	Empty Z values	공간객체의 정점(Vertex)에 Z값이 누락된 오류 (3차원 공간정보일 경우 진단)		점,선,면
	1105	Incorrect ring ordering	Polygon 입력 방향 규칙을 위반한 오류 (바깥Polygon:시계방향, 내부Polygon:반시계방향)		면
	1106	Incorrect segment orientation	각각의 Segment 의 방향이 일정하지 않은 오류. (Shape 파일로 최초 생성된 데이터의 진단 . DB로 구축된 경우는 진단제외)		선,면
	1107	Mismatched attributes	라인 세그먼트 끝점의 Z 또는 M 값이 연결되는 세그먼트 끝점의 Z또는 M 값과 불일치한 오류		선
	1108	Null geometry	공간데이터에 Geometry 값이 없거나 Shape 필드가 없는 오류		점,선,면
	1109	Self intersections	단일 Polygon이 교차되어 있는 오류		면
	1110	Short segment	허용하는 최소단위 값보다 짧은 세그먼트가 있는 오류		선
	1111	Unclosed rings	Polygon의 시작점과 끝점 불일치한 오류		면
진단 설명	공간객체를 표현하는 점선면의 위상 표현의 정확성만을 진단 (참고)ArcGis의 CheckGeometry 기능을 참조하여 진단				

규칙ID	내용	
	진단대상	비고
1101	Discontinuous parts	<ul style="list-style-type: none"> Part로 생성된 객체가 라인이 끊어지거나 불연속으로 구축된 오류. (Shape 파일로 최초 생성된 데이터의 진단 . DB로 구축된 경우는 진단제외)
	선,면 멀티	
1102	Duplicate vertex	<ul style="list-style-type: none"> 동일한 좌표에 두 개 이상의 중복 정점(Vertex)을 가지고 있는 오류
	선,면 싱글,멀티	<ul style="list-style-type: none"> 꼭짓점 중 하나를 삭제

그림 34 라인 끊어짐



- 멀티 파트일 때 한 노선의 도로를 끊어지게 생성되는 오류가 발생

그림 35 정점 중복 예시



- 선이나 다각형에 같은 위치에 있거나 서로 매우 가까운 꼭짓점이 있을 수 있다. 진단에서 지정된 공차가 적당히 작으면 해당 안의 범위에 존재하는 정점들이 결과로 반환됨

I

II

III

IV

데이터 품질진단 방법론

V

규칙ID	내용	
	진단대상	비고
1103	Empty parts	<ul style="list-style-type: none"> 공간정보가 멀티 파트의 경우 그 중 파트 일부가 비어 있는 오류 (Shape 파일만 해당)
	점, 선, 면	<div>멀티</div> <ul style="list-style-type: none"> null이거나 비어있는 대상은 삭제
	<p>그림 36 멀티파트에 비어있는 파트 (예시)</p> 	
	<ul style="list-style-type: none"> 공간객체(지오메트리에는 여러 객체(멀티파트)들 중 하나는 비어있음 : 빈 부분은 다중 선이나 폴리곤과 같은 멀티파트 내에서 발생 	
1104	Empty Z values	<ul style="list-style-type: none"> 공간객체의 정점(Vertex)에 Z값이 누락된 오류 (3차원 공간정보일 경우 진단)
	점, 선, 면	<div>싱글, 멀티</div> <ul style="list-style-type: none"> Z값 수정 (정확한 값을 알 수 없을 경우 0으로 설정)
	<p>그림 37 Z값 Null (예시)</p> 	
	<ul style="list-style-type: none"> 꼭지점의 지오메트리에서 Z(높이)값이 비어있을 때 발생 	

규칙ID	내용	
	진단대상	비고
1105	Incorrect ring ordering	<ul style="list-style-type: none"> 폴리곤은 하나의 외부 경계와 0 개 이상의 내부 경계로 정의되며 내부 경계가 있는 경우 방향이 정확하여야 함 (외부 고리-시계 방향, 내부 고리-시계 반대 방향)
	면	멀티
		<ul style="list-style-type: none"> 지오메트리가 정확한 링 순서를 갖도록 수정
<p>그림 38 입력방향 오류 (예시)</p> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid gray; padding: 5px; text-align: center;"> <p>내부 링이 1개 있는 폴리곤</p> </div> <div style="border: 1px solid gray; padding: 5px; text-align: center;"> <p>내부 링이 2개 있는 폴리곤</p> </div> </div>		
<ul style="list-style-type: none"> 다각형의 바깥쪽 고리를 걷는다고 보면, 오른쪽의 모든 것은 다각형 내부에 있는 것으로 간주하고 왼쪽의 모든 것은 다각형 외부에 있는 것으로 간주한다. 		
1106	Incorrect segment orientation	<ul style="list-style-type: none"> 각각의 Segment 의 방향이 일정하지 않은 오류
	선, 면	싱글
		<ul style="list-style-type: none"> 정확한 세그먼트 방향을 갖도록 지오메트리를 수정
<p>그림 39 Segment 방향 오류 (예시)</p> <div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid gray; padding: 5px; text-align: center;"> <p>오류</p> </div> <div style="font-size: 2em;">➡</div> <div style="border: 1px solid gray; padding: 5px; text-align: center;"> <p>수정</p> </div> </div>		
<ul style="list-style-type: none"> 개별 세그먼트의 방향이 같아야하는데 세그먼트 중 일부가 다른 방향으로 구성된 경우에 발생 (Shape 파일로 최초 생성된 데이터의 진단 . DB로 구축된 경우는 진단제외) 		

I

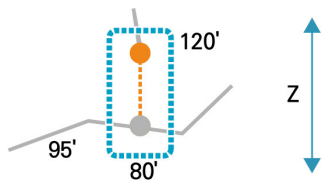
II

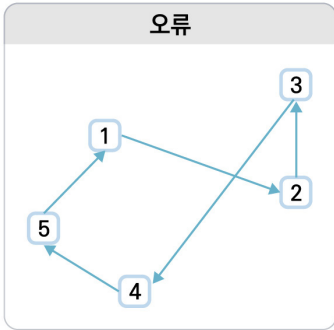
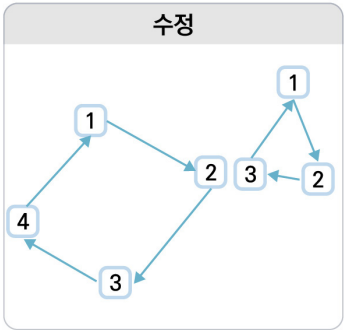
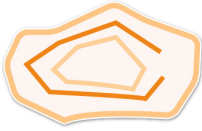
III

IV

데이터 품질진단 방법

V

규칙ID	내용													
	진단대상	비고												
1107	Mismatched attributes													
	라인 세그먼트 끝점의 Z 또는 M 값이 연결되는 세그먼트 끝점의 Z 또는 M 값과 불일치한 오류.													
	선	멀티												
Z값 (좌표)이 일치(연결)되도록 수정														
<div>그림 40 일치하지 않는 속성 (예시)</div> 														
<ul style="list-style-type: none">서로 다른 두 피처가 x, y 평면에서 교차할 때 두 개의 다른 z-값을 가질 수 있어 교차점이 지정된 공차보다 큰 z-값 차이가 있는 교차 폴리 라인을 검사등고선과 지면의 높이를 측정할 수 있고 다리가 지면보다 높을 수 있으므로 등고선과 교량에는 서로 다른 두 개의 표고가 있을 수 있음. 그러나 교량을 교차하는 도로 피처는 교차로에서 동일한 z-값을 가져야 함														
1108	Null geometry													
	공간데이터에 지오메트리 값이 없거나 Shape 필드가 없음													
	점, 선, 면	싱글, 멀티												
클래스에서 지오메트리 값이 없는 피처 삭제														
<div>그림 41 비어있는 지오메트리 (예시)</div> <table><tr><th>OBJECTID</th><th>SHAPE</th><th>FCSUBTYPE</th></tr><tr><td>01</td><td>POLYLINE</td><td>AQ040-BRIDGE LINE</td></tr><tr><td>02</td><td>POLYLINE</td><td>AQ040-BRIDGE LINE</td></tr><tr><td>03</td><td>NUL</td><td>AQ040-BRIDGE LINE</td></tr></table>			OBJECTID	SHAPE	FCSUBTYPE	01	POLYLINE	AQ040-BRIDGE LINE	02	POLYLINE	AQ040-BRIDGE LINE	03	NUL	AQ040-BRIDGE LINE
OBJECTID	SHAPE	FCSUBTYPE												
01	POLYLINE	AQ040-BRIDGE LINE												
02	POLYLINE	AQ040-BRIDGE LINE												
03	NUL	AQ040-BRIDGE LINE												
<ul style="list-style-type: none">형상의 지오메트리가 없거나, 원래 초기화 상태를 벗어나 SHAPE_LEN 및 SHAPE_AREA 필드로 표시되는 기하학적 정보를 포함하지 않을 때 발생														

규칙ID	내용	
	진단대상	비고
1109	Self-intersections	<ul style="list-style-type: none"> 폴리곤의 세그먼트는 서로 교차하지 않아야 함 (단일Polygon이 교차되어 있는 오류)
	선	싱글
	<p>교차하는 segment는 교차점에서 분할하거나 올바른 방향으로 수정</p> <p>그림 42 세그먼트 교차 (예시)</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid gray; padding: 10px; text-align: center;"> <p>오류</p>  </div> <div style="font-size: 2em; margin: 0 10px;">➡</div> <div style="border: 1px solid gray; padding: 10px; text-align: center;"> <p>수정</p>  </div> </div> <ul style="list-style-type: none"> 폴리곤의 세그먼트가 서로 교차되지 않아야 한다. 폴리곤의 경계가 서로 교차하는 모습은 '나비모양' 또는 '8자 모양'의 폴리곤으로 나타난다. 	
1110	Short segment	<ul style="list-style-type: none"> 허용하는 최소단위 값보다 짧은 segment가 있는 경우 (최소단위 이하의 짧은 선형)
	선	싱글
	<p>그림 43 짧은 세그먼트 (예시)</p>  <ul style="list-style-type: none"> 공간객체(도로)의 길이에 비해 허용 기본보다 작은 세그먼트를 찾음. (1미터 미만의 거리) 	

I

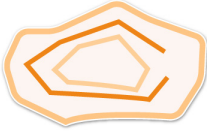
II

III

IV

데이터 품질진단 방법

V

규칙ID	내용	
	진단대상	비고
1111	Unclosed rings	<ul style="list-style-type: none"> 폴리곤의 마지막 세그먼트는 첫 번째 세그먼트의 포인트에 연결되어야 함
	면 싱글	<ul style="list-style-type: none"> 닫히지 않은 링의 끝점과 시작점을 연결
	<p>그림 44 시작점 끝점 불일치(예시)</p> 	
	<ul style="list-style-type: none"> 다각형은 서로 연결된 기하학적 구조인 링 형태로 구성된다. 다각형에 존재할 수 있는 링(ring)은 닫힌 구조를 뜻하므로 링(ring)의 시작점과 끝점이 연결되어 있지 않으면 다각형은 닫히지 않는 모습을 갖는다. 	

▶ 개념적 일관성

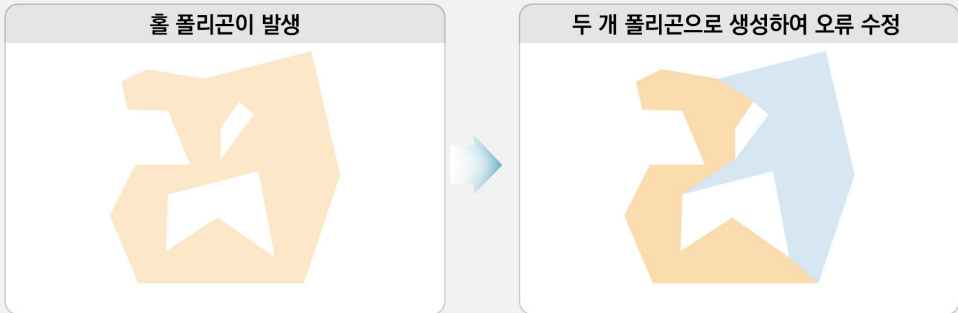
표 36 개념적 일관성 (1/2)

지표	일관성	세부지표	개념적 일관성	진단규칙	개별공간객체 완결성
세부지표정의	<ul style="list-style-type: none"> 공간객체간의 위상관계⁷⁾ 규칙을 준수하고 있는지를 진단하는 지표 				
진단 대상	품질진단 대상 공간정보				면
진단 기법	구조화 기반 공간 진단				프로파일 진단
진단 조건	-				
진단 규칙	<ul style="list-style-type: none"> 공개별 공간정보(레이어) 내의 공간객체들이 공간적 완결성(형식, 관계)을 유지하고 있는지에 대하여 진단 공간정보에 홀 폴리곤(hole polygon) 및 멀티 폴리곤(multi polygon)으로 구성된 공간객체가 영역의 겹침이 없어야 한다. 				
진단 설명	<ul style="list-style-type: none"> 형식) 개별 공간파일에는 같은 유형의 공간정보로만 구성되었는지 진단 관계) 개별 공간파일 내의 객체가 겹침이 있는지 진단 관계) 개별 공간파일 내의 동일 객체가 있는지 진단 (참고) ArcGis의 topology rule기능을 참조하여 진단 - Must not overlap 				

7) 위상관계 : 지도요소들 간의 공간적 관계

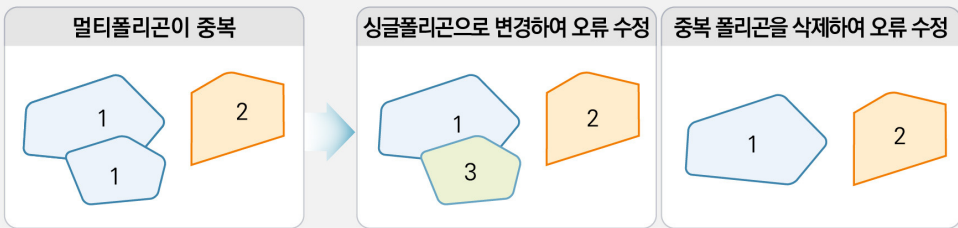
▶ 개별공간객체 완결성 > 홀 폴리곤(hole polygon) 예시

그림 45 개념적 일관성 (예시 1)



▶ 개별공간객체 완결성 > 멀티 폴리곤 객체가 서로 일부 영역 중첩 예시

그림 46 개념적 일관성 (예시 2)



I

II

III

IV

데이터 품질진단 방법

V

▶ 개념적 일관성

표 37 개념적 일관성 (2/2)

지표	일관성	세부지표	개념적 일관성	진단규칙	개념적 일관성
세부지표정의	공간객체간의 위상관계 규칙을 준수하고 있는지를 진단하는 지표				
진단 대상	품질진단 대상 공간정보, 기준공간정보(수치지형도)				선, 면
진단 기법	레이어(layer)기반 공간 진단 (2개 레이어 기반 진단)				업무규칙 진단
진단 조건	작업규정1)에 대한 정의가 요구되는 공간정보에 국한하여 진단				
진단 규칙	공간정보가 개념적 스키마 규칙(공간금칙규칙)에 위배된 공간객체가 없어야 한다.				
진단 설명	<ul style="list-style-type: none"> 기준공간정보와 생성한 공간정보의 레이어를 오버랩을 통하여 사전에 정의한 공간금칙규칙에 위배되는 객체가 없어야 한다. 입력된 두 SHAPE 레이어(layer)에 대해서 서로 오버랩을 통하여 오류 발생 확인 <ul style="list-style-type: none"> 예1) 도로명주소기본도 및 수치지형도와 비교하여 생성한 건물의 위치를 확인 예2) 철도 레이어에 건물의 존재, 호수와 겹쳐지거나 포함된 건물 존재 <p>(참고) ArcGis의 topology rule기능을 참조하여 진단</p> <ul style="list-style-type: none"> - Must be larger than cluster tolerance - Must not overlap with - Must be covered by 				

▶ 개념적 일관성 진단 예시

그림 47 개념적 일관성 (예시 3)

오류로 예상되는 주요 유형별 사례

구분	관로 또는 구조물 관통	인접 관로간의 관경 차이	관로 간 미연결	관로와 맨홀·밸브 미연결
예시 유형				
원인	<ul style="list-style-type: none"> • 깊이 오류 • 관경 오류 	<ul style="list-style-type: none"> • 관경 속성 오류 • 관로 연장 오류 • 관로 위치 오류 	<ul style="list-style-type: none"> • 관로 연장 오류 • 관로 위치 오류 	<ul style="list-style-type: none"> • 연결부 공간정보 미구축 • 맨홀 밸브 오류 • 관로 위치 오류

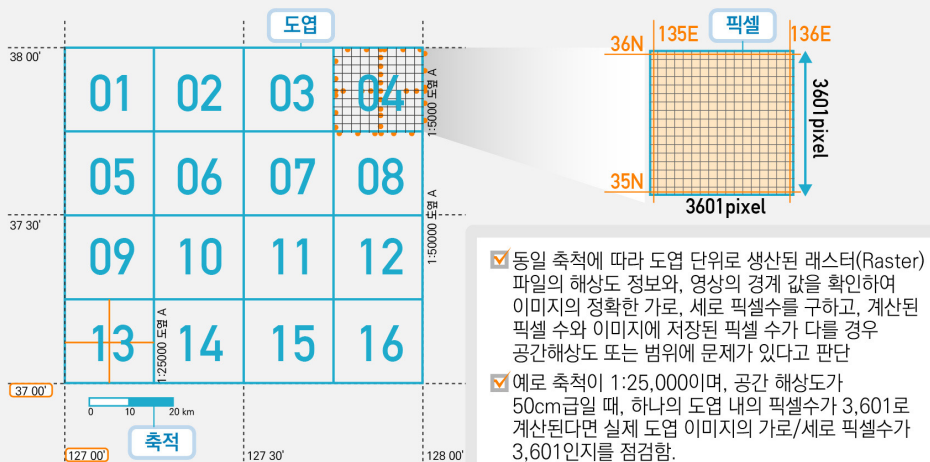
▶ 그리드 유효성

표 38 그리드 유효성 (1/3)

지표	유효성	세부지표	그리드 유효성	진단규칙	위치 정확성 확보
세부지표정의	공간정보의 공간적 위치(좌표계(타원체/투영체계), 도엽 ⁸⁾ , 격자 ⁹⁾)가 일치하는지를 진단하는 지표				
진단 대상	래스터 공간정보(정사영상 ¹⁰⁾ , 수치표고모형 ¹¹⁾ 등)				래스터(Raster) 데이터
진단 기법	단위 영역별로 생성된 표본 추출하여 육안 검사				업무규칙 진단
진단 조건	-				
진단 규칙	래스터(Raster) 공간정보가 도엽 기준과 공간해상도 ¹²⁾ 에 따라 생성된 래스터 이미지의 픽셀수가 일치하여야 한다.				
진단 설명	<ul style="list-style-type: none"> 래스터(Raster) 파일의 공간해상도 정보와 영상의 경계 값을 확인하여 이미지의 가로, 세로 픽셀수를 구하고, 계산된 픽셀 수와 이미지 속성정보의 픽셀 수를 비교하여 래스터(Raster) 파일의 공간해상도의 정확도를 진단 기준에 의한 적정 픽셀 수(가로, 세로 픽셀 수)와 영상 픽셀 수(가로, 세로 픽셀 수)를 비교하여 일치하는가를 확인 예) 영상 양끝의 좌표로부터 계산한 거리와 픽셀수*해상도로 계산한 거리의 일치성 비교 				

▶ 위치 정확성 확보 > 도엽의 위치 정확성 예시

그림 48 그리드 유효성 (예시 1)



- 8) 도엽 : 일정한 도식과 분할되어 한정된 방향으로 통일된 지역을 자른 도면 형식의 지도 1매
- 9) 격자 : 일반적으로 래스터 자료 체계라고도 한다. 래스터 자료를 셀 단위로 저장하는 X, Y좌표 격자망
- 10) 정사영상 : 지표면의 비고(比高)에 의하여 발생하는 사진 상의 각 점의 왜곡을 보정하여 동일 축척이 되도록 지도처럼 만든 영상
- 11) 수치표고모형(Digital Elevation Model : DEM), 적당한 밀도로 분포하는 지점들의 위치 및 표고의 값을 수치화한 후 그 값을 이용하여 지표면의 고도를 수치적으로 표현하는 모형
- 12) 공간해상도(spatial resolution) : 25cm 해상도는 피사체의 25cm를 픽셀(점) 하나로 인식하는 것을 의미함

▶ 그리드 유효성

표 39 그리드 유효성 (2/3)

지표	유효성	세부지표	그리드 유효성	진단규칙	범위 유효성
세부지표정의	• 공간정보의 공간적 위치(격자, 좌표계(타원체/투영체계))가 일치하는지를 진단하는 지표				
진단 대상	• 래스터(Raster) 공간정보(도엽)				래스터(Raster) 데이터
진단 기법	• 단위 영역별로 생성된 표본 추출하여 육안 검사				업무규칙 진단
진단 조건	-				
진단 규칙	• 도엽으로 이루어진 래스터(Raster) 공간정보의 공간적 범위가 도엽 및 좌표계의 범위로부터 유효한지 여부를 진단				
진단 설명	• 정의된 축척의 공간적 범위 기준에 따른 도엽의 범위(최소값(min), 최대값(max))가 유효한가를 진단 • 정의된 표준 도엽 안에 정확히 공간정보가 들어오는지 검사				

▶ 범위 유효성 진단 > 격자 간격 정확도 예시

예시 화면과 같이 서대문구를 도엽으로 생산시 사전에 정의된 축척별 도곽¹³⁾의 공간적 범위 기준에 의하여 도곽의 최소값(min), 최대값(max)을 검사하여 도곽의 범위(최소값(min), 최대값(max))가 유효한가를 진단

그림 49 그리드 유효성 (예시 2)



13) 도곽(圖廓) : 일정한 크기에 따라 분할된 지도의 가장자리에 그려진 경계선

▶ 그리드 유효성

표 40 그리드 유효성 (3/3)

지표	유효성	세부지표	그리드 유효성	진단규칙	공간참조체계 정확성
세부지표정의	• 공간정보의 공간적 위치(격자, 좌표계(타원체/투영체계))가 일치하는지를 진단하는 지표				
진단 대상	• 공간정보(점,선,면)				래스터(Raster) 데이터
진단 기법	• 레이어(layer) 기반 공간 진단(단일 레이어 기반 진단)				업무규칙 진단
진단 조건	-				
진단 규칙	• 좌표계(타원체/투영체계)가 가지는 공간적 범위에 공간객체들이 해당 공간범위 내에 있어야 한다.				
진단 설명	<ul style="list-style-type: none"> • prj 파일에 정의된 좌표계(타원체/투영체계) 및 공간범위의 확인을 통하여 레이어(layer)에 속한 객체들의 좌표(최소값(min)과 최대값(max))이 주어진 공간적 범위를 벗어났는지를 진단 (타 공간정보와 융합 시 부정확한 좌표계 등으로 인하여 위치 오류가 발생함) • SHAPE 데이터를 읽어서 TM중부 GRS80 20만, 60만 범위를 벗어났는지를 확인 (예) 'GIS건물통합' 레이어(layer)는 Korean_1985의 Korea_Central_Belt 좌표계를 사용하고 있으므로 X 좌표는 107,516.3257~292,483.6743, Y 좌표는 455,569.7688~537,099.1393의 범위 내에 있는지 확인 				

▶ 그리드 유효성 > 공간참조체계 정확성 진단 예시

공간정보가 동일 좌표계(타원체/투영체계) 기준으로 공간정보 범위 이탈 사례

그림 50 공간참조체계 정확성 진단 (예시)

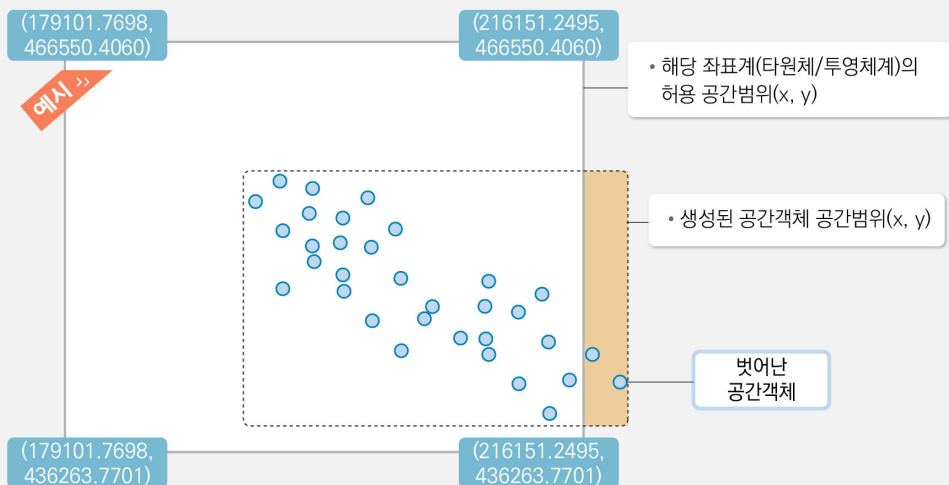


그림 51 그리드 유효성 (예시 3)

▶ 범위 유효성 진단 > 공간정보들의 좌표계(타원체/투영체계) 불일치 사례

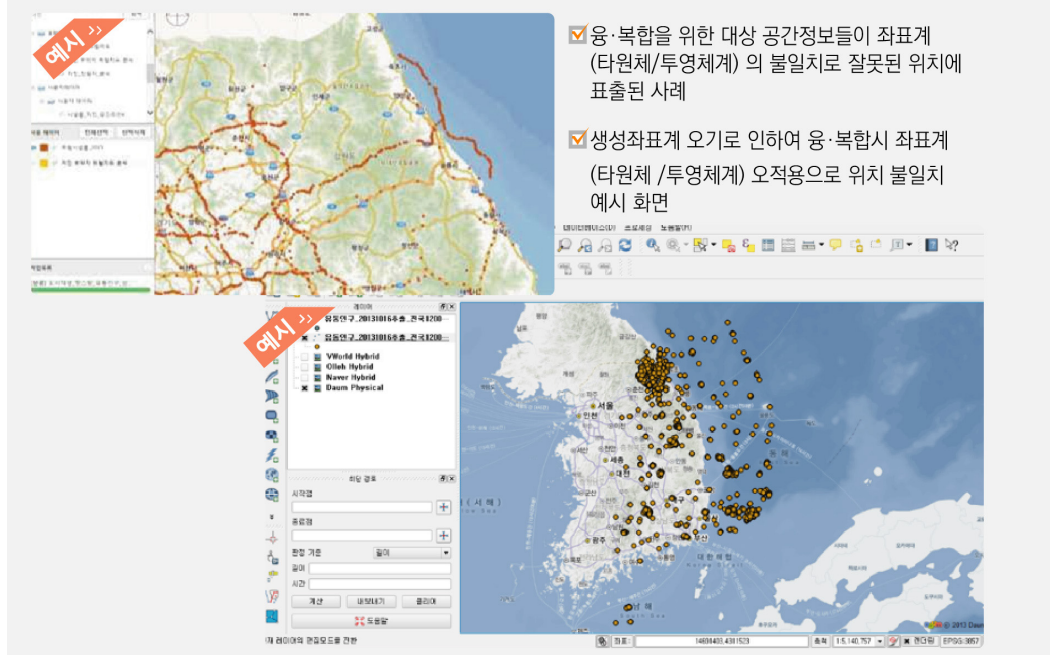
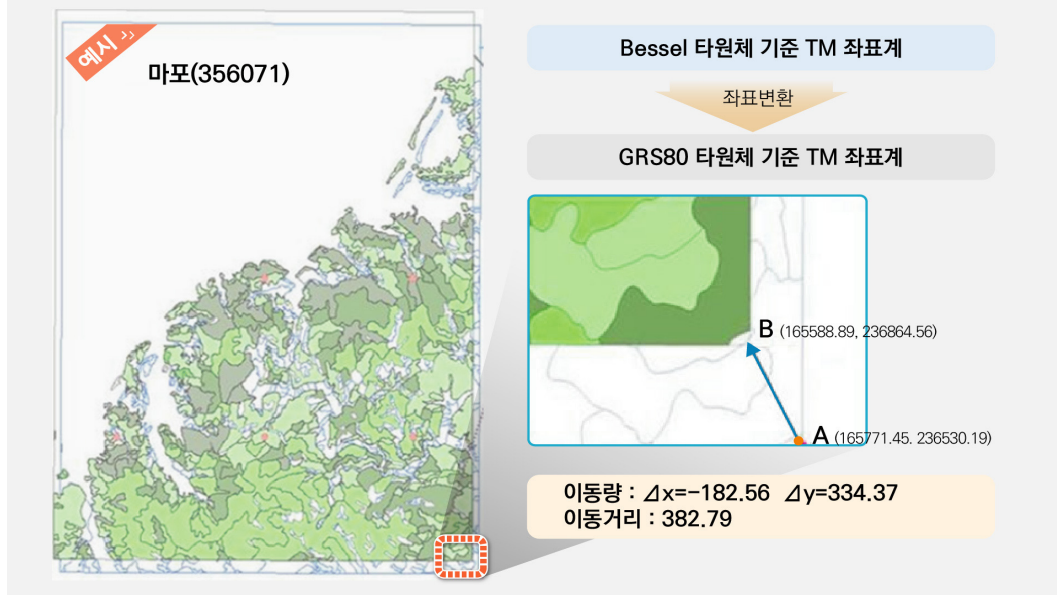


그림 52 그리드 유효성 (예시 4)

▶ 범위 유효성 진단 › 다른 타원체 사용으로 인한 공간정보 불일치 사례

- 공간정보의 구축 및 관리 등에 관한 법률 시행령 [별표 2] <개정 2015.6.1.> 직각좌표의 기준(제7조제3항 관련)
- PCS_ITRF2000_2060_TM.prj : 현 세계측지계 가산수치 N60만 E20만 (2010년 이후의 세계측지계)



3.4 비정형 데이터 진단규칙

3.4.1 비정형 데이터의 구성 및 진단 범위

비정형 데이터는 문서, 이미지, 음성, 영상 등과 같이 비구조 형태를 가지고 있는 데이터로 비구조 형태의 객체 데이터와 객체 데이터에 대한 객체 메타 데이터로 구성되며, 비즈니스 목적에 따라 관리하는 관리정보를 갖고 있을 수 있다.

그림 53 비정형 데이터 구성 및 진단 대상

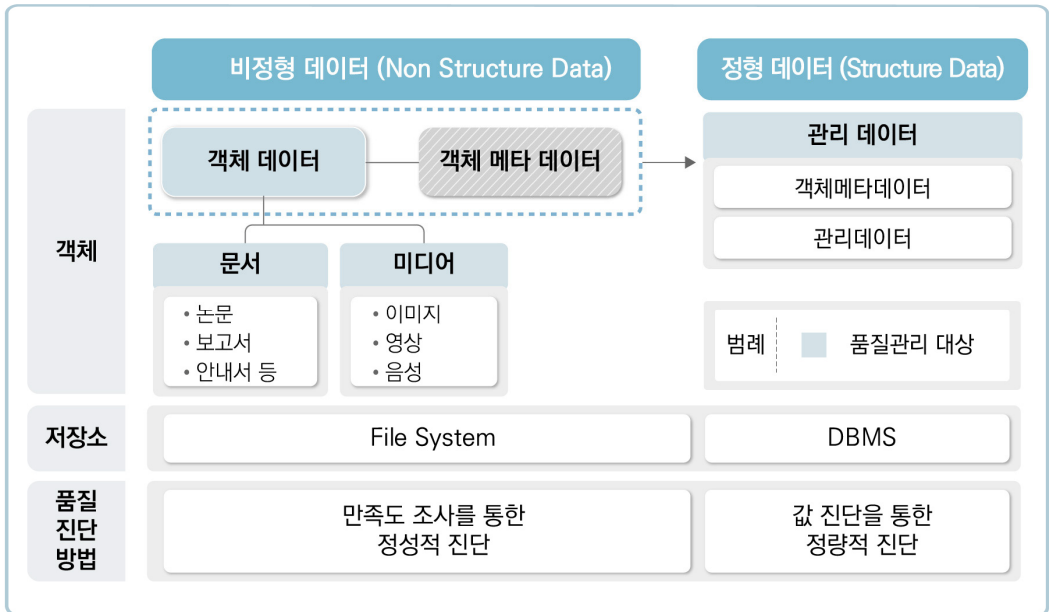


표 41 비정형 데이터 구성요소 설명

대상 구분	설 명	예
객체 데이터	• 비정형 구조 형태의 데이터로 문서파일, 미디어 파일 자체	• 문서, 이미지, 음성, 영상 등
객체메타데이터	• 객체정보를 생성하면서 자동으로 생성되는 정보	• 파일명, 파일형식, 해상도, 생성일자 등
관리 데이터	• 활용을 위해 객체정보를 설명하는 관리데이터 항목 및 객체메타데이터 항목 등의 데이터로 구성된 정보	• 저장위치, 주제영역, 파일명, 파일형식 등
문서	• 문서 작성기로 생성한 문서 정보(고유의 저장형식을 가짐)	• txt, hwp, doc, pdf 등
미디어	• 이미지, 음성, 영상 등의 미디어 정보(고유의 저장형식을 가짐)	• jpg, gif, png, mp3, mp4 등

비정형 이미지 데이터인 “주차장”에 대한 구성은 비정형 데이터인 “주차장 이미지 파일”, 이미지 파일에 대한 메타정보, 주차장 이미지를 관리하는 필요한 “주차장 메타정보”와 주차장관리 업무수행에 필요한 “주차장관리정보”로 구성된 예시는 다음과 같다.

그림 54 비정형 데이터의 구성 예시

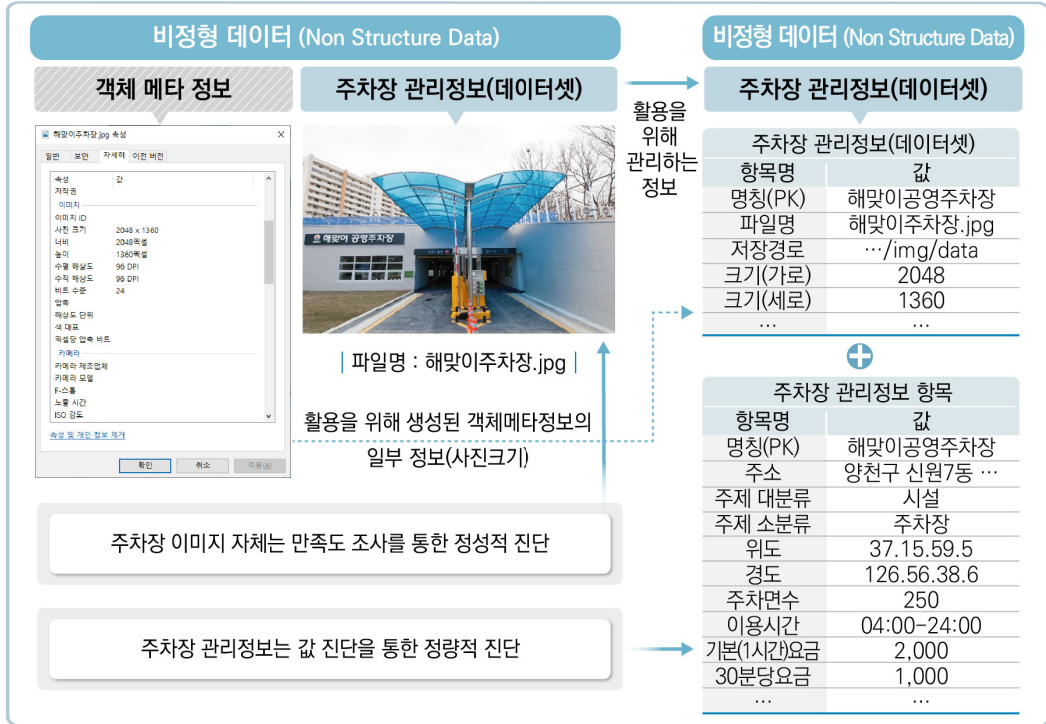
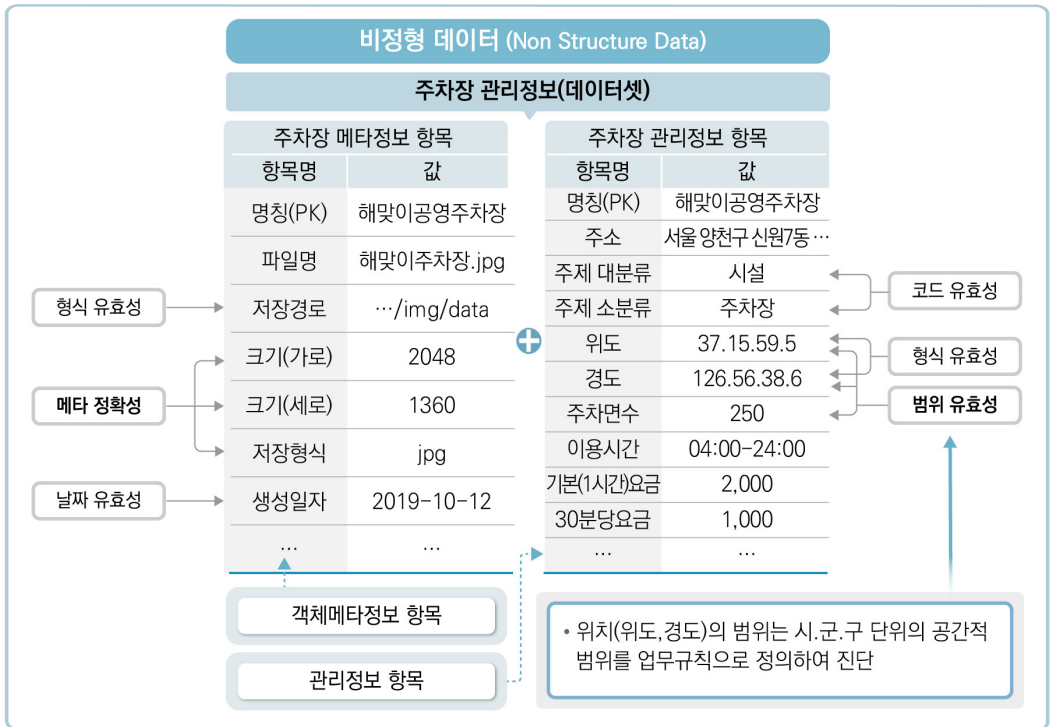


그림 55 비정형 데이터의 진단방법 예시



3.4.2 비정형 데이터의 진단 방향

비정형 데이터 자체에 대한 정량적인 품질진단은 측정의 기준 및 도구의 부재로 현실적인 진단이 어려우므로, 만족도 조사 등을 통하여 정성적 진단이 필요하며, 향후 인공지능 기술을 적용한 진단 도구의 마련이 필요하다.

▶ 문서 데이터의 품질진단 방향

문서 데이터는 활용에 필요한 정보를 생성하기 위해서는 텍스트 마이닝 기술을 이용하여 활용에 필요한 정보를 추출하여 관리하여야 한다.

▶ 이미지 데이터의 품질진단 방향

이미지 데이터는 활용 목적을 고려하여 미디어의 유형별로 품질에 영향을 주는 관리 가능한 요소를 선정하며, 요소별 기준을 정의하여 관리하여야 한다.

이미지 데이터의 인공지능 기술을 활용하여 이미지의 내용에 대한 태깅의 자동화가 필요하다.

I

II

III

IV

데이터 품질진단 방법

V

04

품질진단 실시



“품질진단 규칙 정의”에서 도출된 진단 스크립트(SQL)를 실행하여 품질진단을 실시하고 진단결과 오류에 대해 업무담당자의 검토 후 재 정의가 필요한 경우 진단규칙을 재정의하여 진단규칙을 확정하며, 진단규칙과 진단Script 3종은 재사용이 가능하도록 자산화하여 체계적으로 관리한다.

4.1 품질진단 실시

아래 예시는 품질진단 대상에 대해 프로파일을 수행하고 그에 따른 결과 분석을 통해 진단규칙을 도출하는 예시이다.

▶ 여부 유효성 진단 실시 예시

1. 진단 실시 대상

고객테이블의 통합고객여부 항목(컬럼) 진단 실시 예시

2. 프로파일링 분석

2.1 항목(컬럼) 분포 값 조회(DISTINCT)

- 유효 값을 파악할 수 없는 경우 항목(컬럼) 분석을 통해 진단규칙 및 오류 예측이 가능(프로파일링)
- 항목(컬럼)을 구성하는 2개의 유효 값 이외의 데이터를 오류로 판단

예) 유효값 : Y/N, T/F, 0/1, 1/2

데이터셋 (테이블)명	항목(컬럼)명	항목(컬럼)값	건수
고객	통합고객여부	Y	90
		N	10
		1	2
		0	1

3. 프로파일 결과 분석 및 진단규칙 도출

- 통합고객여부 항목(컬럼)의 데이터 분포를 통해 'Y','N'값 이외의 데이터를 오류로 판단

(진단규칙 SQL - Oracle 11g 기준)

■ 전체건수

```
SELECT COUNT(*) AS 전체건수
FROM 고객;
```

■ 오류건수 추출 SQL

```
SELECT 통합고객여부
, COUNT(*) AS 오류건수
FROM 고객
WHERE 통합고객여부 NOT IN ('Y', 'N')
GROUP BY 통합고객여부;
```

■ 오류 추출 SQL

```
SELECT  고객번호, 통합고객여부
FROM    고객
WHERE   통합고객여부 NOT IN ( 'Y' , 'N' );
```

4.2 품질진단 실시 결과 검토

품질진단을 수행한 후 올바른 진단규칙으로 설정되었는지에 대한 검증을 실시한다. 아래 예시는 품질진단을 수행 하였지만 진단규칙이 올바르게 정의 되어 있지 않은 경우의 예시이다.

▶ 품질진단 실시 결과 검토 예시

1. 진단 실시 대상

- 고객테이블의 통합고객여부 항목(컬럼) 진단 실시 예시

2. 진단 기준 및 결과

2.1 메타정보(도메인 정의서)

- 메타정보인 도메인정의를 기준으로 유효 값 0, 1 값 이외에는 오류로 검출되는 진단규칙 도출
- 항목(컬럼)을 구성하는 2개의 유효 값('0','1') 이외의 데이터를 오류로 판단

도메인명	TYPE	길이	정의
여부	VARCHAR2	1	유효값은 '0' , '1' 로 정의한다.

(진단규칙 SQL - Oracle 11g 기준)

```
SELECT  통합고객여부
        , COUNT(*) AS 오류건수
FROM    고객
WHERE   통합고객여부 NOT IN ( '1' , '0' )
GROUP BY 통합고객여부;
```

2.2 진단 결과 조회

- 진단규칙 통합고객여부 항목(컬럼)의 유효 값 '0','1'을 기준으로 진단
- 실시 결과 오류율 약 95%로 도출

진단규칙명	항목(컬럼)값	건수	오류율
여부-0/1	Y	70	약 98%
	N	150	
	0	4	

(조회 SQL - Oracle 11g 기준)

```
SELECT  통합고객여부 AS 항목(컬럼)값
        , COUNT(*) AS 오류건수
FROM    고객
GROUP BY 통합고객여부;
```

3. 진단기준 및 결과 검토 후 진단규칙 재정의

- 진단기준 및 진단 결과 검토 결과 유효 값 재설정('Y','N')을 통한 재진단 실시 필요

(진단규칙 SQL - Oracle 11g 기준)

■ 재정의 오류건수 추출 SQL

```
SELECT  통합고객여부
        , COUNT(*) AS 오류건수
FROM    고객
WHERE   통합고객여부 NOT IN ( 'Y' , 'N' )
GROUP BY 통합고객여부;
```

05

품질진단 결과



최종 오류 확정 후 품질진단 유형, 품질진단대상 내역, 품질진단 규칙 정의내역, 품질진단 결과내역 (값, 구조, 표준, 관리체계) 등을 포함하여 “품질진단 결과서”를 작성한다. 품질관리 담당자는 “품질진단 결과서”를 관련 이해관계자들에게 공유한다.

표 42 품질진단 결과서 구성

항 목	내 용
품질진단 수행 현황	<ul style="list-style-type: none"> 품질진단 대상의 시스템 정보, 진단 대상(계획), 진단 대상(수행)에 대하여 기술한다. 예) 저장영역, DBMS, 주제영역, 진단대상(계획) 테이블 및 컬럼 수, 진단대상(수행) 테이블 및 컬럼 수, 미수행 사유
품질진단 결과내역	<ul style="list-style-type: none"> 품질진단 결과내역을 요약 기술한다. <ol style="list-style-type: none"> 데이터 저장영역 별 오류현황 <ul style="list-style-type: none"> 정량진단(값, 구조, 표준) 지표 별 오류 현황 <ul style="list-style-type: none"> 품질지표 별 오류율(정량진단, 정성진단) 내용 지표 별 상세 오류 현황 <ul style="list-style-type: none"> 지표 별 테이블 수, 컬럼 수, 전체건수, 오류건수, 오류율 내용
품질진단결과 총평	<ul style="list-style-type: none"> 품질진단결과 이후 진행되어야 할 사항에 대한 간단한 요약 및 일정을 기술한다. 예) 품질진단 결과의 전반적인 오류율을 비교하였을 때 유효성 지표에 대한 오류율이 높아 향후 유효성 지표 관련 품질진단 대상의 지속적인 관리가 필요해 보임. 품질진단 계획대비 미수행 1개 테이블 10개 컬럼에 대한 진단실시는 2차 품질진단에 적용하여 수행 예정

01

개요



데이터 품질진단 결과의 데이터 값 오류에 대한 근본원인을 분석하여 전체 시스템 운영 관점에서 오류 유형별 개선기회를 도출한다.

그림 56 진단결과 분석(Analyze) 절차



표 43 진단결과 분석(Analyze) 절차 설명

단계	내 용
A1.오류원인 분석	<ul style="list-style-type: none"> 데이터의 부정확한 값이 발생되고 유입되는 오류 유형별 오류 원인을 분석 하고 오류 원인 분석 결과서를 작성한다. - 오류원인유형 : 값, 구조, 표준, 관리체계, 응용프로그램
A2.개선 기회 도출	<ul style="list-style-type: none"> 데이터 값 오류 발생 근본원인을 제거하기 위하여 전체 시스템 운영 관점에서 오류 유형별 개선 기회를 도출 하고 진단결과 분석서를 작성한다. - 산출물 : [서식-9] 오류원인 분석 내역, [서식-10]품질진단 결과 분석서

I

II

III

IV

데이터 품질진단 방법

V

02

오류원인 분석



“제3장 품질진단 수행(Measure)”에서 최종 오류로 확정 된 데이터에 대하여 오류발생 원인을 파악한다. 오류의 원인은 단일 혹은 복합적인 요소가 존재 할 수 있으며, 반복적인 오류 발생을 막기 위해서는 근본적인 오류 데이터 발생 원인을 찾아내는 것이 중요하다.

오류 데이터 발생 원인은 다양한 요소들(값, 구조, 표준, 관리체계, 응용프로그램)이 존재하기 때문에 오류 방지를 위한 관련업무 이해관계자들의 협조 및 노력이 필요하다.

▶ 유형별 오류 원인 (예시)

표 44 유형별 오류원인 (예시)

유형	오류 원인
값 오류	<ul style="list-style-type: none"> 차세대 구축 시 AS_IS 데이터를 TO_BE 데이터로 이행 작업 시 오류 사용자 입력 오류
구조 오류	<ul style="list-style-type: none"> 시스템 구축 이후 변경관리 미흡 데이터 모델관리의 가이드 지침서 적용의 부재
표준 오류	<ul style="list-style-type: none"> 시스템 구축 초기의 표준은 존재하나 향후 변경관리가 미흡 표준단어·표준용어·표준도메인·표준코드 정의 미흡으로 정비필요
관리체계 오류	<ul style="list-style-type: none"> 품질관리 지침, 정책 등의 부재 품질관리 프로세스 미흡, 조직 및 역량 부족 데이터 오너십 미정의
응용프로그램 오류	<ul style="list-style-type: none"> 응용프로그램의 입력값 검증 기능 부재(응용프로그램) 연계데이터 정합성 검증 미흡으로 오류

03

개선기회 도출



분석한 오류의 근본 원인을 제거하기 위해서 구체적인 개선 기회를 도출한다.

오류 데이터 값뿐만 아니라, 데이터 값 오류의 근본원인이 되는 구조, 표준, 관리체계, 응용프로그램도 개선 대상에 포함되어야 한다.

▶ 개선 기회 유형 (예시)

표 45 개선기회 유형 (예시)

유형	오류 원인
값	<ul style="list-style-type: none"> “고객”테이블의 “고객번호” 정제, “주문”테이블의 “주문일”과 “배송시작일”간의 선·후행 관계 정제 등
구조	<ul style="list-style-type: none"> “고객”테이블 정규화 필요, “주문” 테이블 식별자 항목 조정, “회원관리·고객”테이블의 데이터 구조 통합 등
표준	<ul style="list-style-type: none"> 표준 단어, 용어, 도메인 관리 미흡으로 인한 단어, 용어, 도메인의 표준 정비 수행 등
관리체계	<ul style="list-style-type: none"> 오너십 관리 부재로 인한 모델·데이터베이스 변경·통제 프로세스 수립, 데이터 표준에 대한 변경·통제 프로세스 및 조직 등
응용프로그램	<ul style="list-style-type: none"> “고객정보 등록”프로그램의 “고객 정보” 검증 프로세스 보완 “주문배송등록”프로그램의 “주문”과 “배송시작일”간의 선·후행 관계 검증 프로세스 보완 등

오류원인 분석과 개선 기회 도출 과정을 수행 후 품질진단결과 분석서를 작성한다.

품질진단결과 분석서 작성 시 품질진단 대상정보, 품질진단 실시 결과, 오류유형/원인 분석, 개선 기회 유형 등의 항목을 포함하여 작성하고 관련 이해관계자들에게 공유한다.

I

II

III

IV

데이터 품질진단 방법

V

▶ 진단결과 분석서 구성

표 46 진단결과 분석서 구성

유형	내용
품질진단 결과내역	<ul style="list-style-type: none"> 품질진단결과내역을 요약 기술한다. <ol style="list-style-type: none"> 데이터 저장영역 별 오류현황 <ul style="list-style-type: none"> 정량진단(값, 구조, 표준) 지표 별 오류 현황 <ul style="list-style-type: none"> 품질지표 별 오류율(정량진단, 정성진단) 내용 지표 별 상세 오류 현황 <ul style="list-style-type: none"> 지표 별 테이블 수, 컬럼 수, 전체건수, 오류건수, 오류율 내용
오류 원인 현황	<ul style="list-style-type: none"> 품질진단결과와 분석내역의 오류발생유형을 요약 기술한다. <ul style="list-style-type: none"> 저장영역, DBMS 별 오류발생유형 수 요약 (값, 구조, 표준, 관리체계, 응용프로그램)
개선 기회 도출	<ul style="list-style-type: none"> 오류발생 원인에 따른 개선 기회에 대해 기술한다. <ul style="list-style-type: none"> 오류발생유형(값,구조,표준, 관리체계, 응용프로그램) 별 오류발생원인 및 개선 기회 내용

(산출물 : [서식-10]품질진단 결과 분석서)

01

개요



“개선수행(Improve)”은 오류를 유발하는 근본 원인을 제거하기 위하여 품질개선 과제 정의 및 품질개선 계획을 수립하고, 개선 환경을 준비하여, 실제 개선을 수행하는 단계이다.

한정된 시간과 비용, 인력으로 모든 문제를 해결하기 어려울 수 있기 때문에 품질개선 효과를 거두기 위해서는 품질에 영향을 미치는 요소 중 개선 효과가 크고 중요한 문제들을 우선적으로 선정하여 개선한다.

그림 57 개선수행(Improve) 절차



표 47 개선수행(Improve) 절차 설명

단계	내 용
11. 품질개선 과제 정의	<ul style="list-style-type: none"> 품질개선 우선순위 선정기준에 따라 과제수행 우선순위를 결정하고 추진 일정을 정의
12. 품질개선 계획 수립	<ul style="list-style-type: none"> 추진체계 및 일정, 예산 등 품질개선 추진을 위한 구체적인 개선 계획서를 작성 - 산출물 : [서식-11]품질개선 계획서)
13. 품질개선 환경 준비	<ul style="list-style-type: none"> 품질개선의 수행에 앞서 품질개선을 효과적으로 수행하기 위해 사전에 준비해야 할 환경을 준비
14. 품질개선 수행	<ul style="list-style-type: none"> 품질개선 수행계획에 따라 오류의 근본 원인 별로 실제 품질 개선을 수행하는 활동 - 산출물 : [서식-12]품질개선 결과내역, [서식-13]품질개선 결과서

02

품질개선 과제 정의



데이터 오류를 유발하는 근본 원인을 제거하기 위해 시급성 및 중요도 등을 고려하여 개선과제 우선순위를 선정하고 수행 방식을 결정한다.

2.1 우선순위 선정

개선과제에 대해 오류파급효과(↑), 시급성(↑), 중요도(↑), 시스템영향도(↓) 관점을 고려하여 개선과제 우선순위를 선정한다.

개선과제가 개별 수행 가능한 단위이면 그대로 순위를 선정하고, 개선과제 안에서도 단계별·연차별로 수행해야하는 경우가 있으면 과제를 세분화하여 우선순위를 선정한다.

개선과제 사이에도 선·후행 관계가 있을 수 있으므로 이를 잘 파악하여 우선순위를 정의한다.

선·후행 관계가 있는 우선순위 정의 예시

- 표준화 수립과 코드 데이터 보정의 개선을 수행해야하는 경우
- 표준화 수립 → 코드 데이터를 보정
- 응용프로그램과 데이터영역 보정 그리고 업무규정 재정비가 필요 한 경우
- 업무규정 재정비 → 응용프로그램과 데이터 영역 보정

▶ 개선과제 우선순위 선정 기준

표 48 개선과제 우선순위 선정 기준

기준 항목	설명	순위 결정 기준
오류파급효과	개선하지 않을 경우 오류가 미치는 파급 정도	파급효과가 클수록 우선순위가 높다.
시급성	빠른 시일 안에 긴급하게 개선을 요하는 정도	시급성이 높을수록 우선순위가 높다.
중요도	개선과제가 업무적(분석·활용)으로 중요한 정도	중요도가 높을수록 우선순위가 높다.
시스템영향도	개선과제 수행 시 시스템에 미치는 영향 정도	시스템 영향도가 높을수록 우선순위가 낮다.

▶ 개선 우선순위 선정기준에 의한 순위 결정 (예시)

표 49 개선 우선순위 선정기준에 의한 순위 결정 (예시)

개선과제	기준 항목별 평가								개선 순위	
	파급효과(↑)		시급성(↑)		중요도(↑)		시스템영향도(↓)			
기상정보 가공· 융합 프로그램 보완	중	2	상	3	상	3	중	2	10	1
코드 데이터 표준 적용	상	3	중	2	상	3	하	1	9	2
....										

※ 순위결정은 상:3/중:2/하:1 3단계로 평가

※ 기준항목에 가중치를 주어 평가 할 수도 있다.

2.2 개선과제 수행 방식 선정

개선과제 우선순위가 정의되면 시스템 유지보수 동안에 할 것인지 별도의 품질개선 사업에서 할 것인지 시스템구축 사업 동안에 할 것인지 등 개선과제에 대한 수행 방식을 결정한다.

▶ 개선과제 수행 방식 결정 시 고려사항

표 50 개선과제 수행 방식 결정 시 고려사항

유 형	내 용
시스템 유지보수	<ul style="list-style-type: none"> 시스템 유지보수를 하는 동안에 개선과제 수행 개선과제가 데이터 값을 변경하거나, 응용 프로그램 일부가 변경 등 시스템 영향도가 적은 과제를 수행하는 경우 해당
별도의 품질개선 사업	<ul style="list-style-type: none"> 시스템 유지보수 동안에 수행하기에는 범위가 커서 별도의 품질개선 사업 수행 심각한 품질 이슈로 인하여 시스템 영향도가 크고 시급하게 개선이 필요한 경우에 당해 또는 내년에 차세대와 같은 재개발 사업이 없을 때 해당
시스템 (재)구축 사업	<ul style="list-style-type: none"> 차세대 사업 또는 신규 시스템 도입 시 개선과제 수행 개선과제가 프로세스를 변경해야 하거나, 다양한 영역에 걸쳐 있어 해당 영역만 변경하기 어려운 경우에 해당

03

품질개선 계획 수립



품질개선과제의 우선순위와 수행방식이 정의되면, 이를 기반으로 개선을 위한 계획을 수립한다. 품질개선 계획 수립은 사업계획서와 같은 것으로, 사업 추진의 목적, 품질개선 대상, 개선과제 정의, 품질개선 수행 일정 등 품질개선 수행을 위한 구체적인 개선 계획서를 작성하는 것이다.

▶ 품질개선 계획서 항목

표 51 품질개선 계획서 항목

항 목	내 용
품질개선 목적	<ul style="list-style-type: none"> 품질개선 수행 목적을 기술한다. 예) 시스템 품질진단 정기 1차 점검에 따른 오류에 대한 품질 개선을 실시하고자 함
품질개선 대상	<ul style="list-style-type: none"> 품질오류 개선 대상에 대해 기술한다. 1) 저장영역, DBMS 별 오류발생 유형 수 요약 (값, 구조, 표준, 관리체계, 응용프로그램) 2) 저장영역 별 오류 발생 현황 내용(오류가 발생 된 테이블 컬럼의 개수)
개선과제 정의	<ul style="list-style-type: none"> 개선 우선순위 및 개선과제를 정의하여 기술한다. 1) 개선 과제 별 우선순위 선정 <ul style="list-style-type: none"> - 우선순위, 개선 수행 방식, 개선과제 내용 2) 저장영역 별 우선순위 도출 <ul style="list-style-type: none"> - 저장영역, DBMS 별 오류발생유형의 우선순위 내용
품질개선 수행 일정	<ul style="list-style-type: none"> 품질개선 단계별 수행일정을 기술한다. 1) 최우선 과제 <ul style="list-style-type: none"> - 품질개선환경구성, 품질개선 수행(최우선), 품질개선 결과서 작성 2) 단기 및 중장기 과제 <ul style="list-style-type: none"> - 우선순위(단기, 중장기), 과제명, 수행일정, 수행방법, 수행예산 등의 내용

(산출물 : [서식-11]품질개선 계획서)

04

품질개선 환경 준비



품질개선의 수행에 앞서 품질개선을 효과적으로 수행하기 위해 이해당사자간 협업체계를 구성하는 등 사업수행에 필요한 환경을 준비하는 단계다.

▶ 품질개선 환경준비 내역

표 52 품질개선 환경준비 내역

항 목	설 명
협조체계 구성 (외부 기관 연계 시)	<ul style="list-style-type: none"> 외부기관 연계 관점에서 개선이 필요한 경우 외부 연계기관 담당자를 포함한 협업 체계를 구성한다.
관련 산출물 준비	<ul style="list-style-type: none"> 품질개선 수행에 필요한 관련 산출물을 식별하고 현행화된 산출물을 준비한다.
개선 대상 준비	<ul style="list-style-type: none"> 품질개선을 수행 할 데이터 또는 프로그램을 준비한다.

I

II

III

IV

데이터 품질진단 방법론

V

05

품질개선 수행



품질개선과제에 대해 개선활동을 수행하고, 수행결과에 대해서 수행결과보고서를 작성한다.

▶ 개선영역 별 세부 개선 사항

품질개선 수행은 품질 관리체계, 표준화, 값, 구조, 응용프로그램 등 개선영역별로 수행한다.

표 53 개선영역별 세부 개선 사항

개선 영역	설명							
품질관리체계 개선	<ul style="list-style-type: none"> 품질 관리에 대한 규정·지침, 절차, 조직 및 역할, 관리도구 활용 방안, 변화관리 방안 등을 정의하여 지속적인 품질관리를 수행할 수 있는 체계를 수립한다. 							
	<table border="1"> <tr> <td>규정·지침, 절차</td><td> <ul style="list-style-type: none"> 기관의 정보화 비전과 미션을 포함하는 데이터 품질 관리 규정(전사 품질 관리 프로세스), 지침, 절차 </td></tr> <tr> <td>조직 및 역할</td><td> <ul style="list-style-type: none"> 품질 관리에 필요한 조직 구성 및 역할(Role), 미션(Mission), 목표(Goal), 성과평가지표(KPI) 등 </td></tr> <tr> <td>관리도구</td><td> <ul style="list-style-type: none"> 품질 개선 프로젝트 이후 품질 통제를 위한 메타데이터 관리 도구, 품질진단 도구 등의 도입 계획 </td></tr> <tr> <td>변화관리 방안</td><td> <ul style="list-style-type: none"> 품질 관리 규정 및 지침, 조직, 도구 등 신규 도입되는 품질 관리 내용에 대한 변화관리 계획 </td></tr> </table>	규정·지침, 절차	<ul style="list-style-type: none"> 기관의 정보화 비전과 미션을 포함하는 데이터 품질 관리 규정(전사 품질 관리 프로세스), 지침, 절차 	조직 및 역할	<ul style="list-style-type: none"> 품질 관리에 필요한 조직 구성 및 역할(Role), 미션(Mission), 목표(Goal), 성과평가지표(KPI) 등 	관리도구	<ul style="list-style-type: none"> 품질 개선 프로젝트 이후 품질 통제를 위한 메타데이터 관리 도구, 품질진단 도구 등의 도입 계획 	변화관리 방안
규정·지침, 절차	<ul style="list-style-type: none"> 기관의 정보화 비전과 미션을 포함하는 데이터 품질 관리 규정(전사 품질 관리 프로세스), 지침, 절차 							
조직 및 역할	<ul style="list-style-type: none"> 품질 관리에 필요한 조직 구성 및 역할(Role), 미션(Mission), 목표(Goal), 성과평가지표(KPI) 등 							
관리도구	<ul style="list-style-type: none"> 품질 개선 프로젝트 이후 품질 통제를 위한 메타데이터 관리 도구, 품질진단 도구 등의 도입 계획 							
변화관리 방안	<ul style="list-style-type: none"> 품질 관리 규정 및 지침, 조직, 도구 등 신규 도입되는 품질 관리 내용에 대한 변화관리 계획 							
표준화 개선	<ul style="list-style-type: none"> 시스템에 사용되는 코드, 용어, 도메인을 분석하여 공통으로 적용 가능한 표준 코드집, 표준용어 사전, 표준 도메인 사전 등을 조직의 전사 차원에서 구축한다. 표준화 수립 결과는 신규 시스템 구축일 경우 일괄 적용이 가능하나, 운영 중인 시스템의 경우는 장기적인 계획을 가지고 쉽게 반영할 수 있는 시스템 영역부터 보완해 나가도록 한다. 							
값 보정	<ul style="list-style-type: none"> 오류 데이터의 값을 직접 보정한다. 							
구조 보정	<ul style="list-style-type: none"> 오류 데이터를 발생시키는 구조적 결함을 보정한다. 							
응용프로그램 보정	<ul style="list-style-type: none"> 오류 데이터 유입 방지를 위한 오류의 원인이 되는 응용프로그램을 보정한다. 							

▶ 수행 결과 보고서 작성

- 품질관리체계 개선, 표준화 개선, 값·구조·응용프로그램 보정 등의 품질개선 수행이 끝나면 개선 수행 결과를 보고서 형태로 작성
- 오류현황 및 수행 결과는 개선 역별(품질관리체계, 표준화, 값, 구조, 응용프로그램)로 작성

▶ 품질개선 결과서 구성

표 54 품질개선 결과서 내역

항목	내용
품질 개선 대상	<ul style="list-style-type: none"> • 품질오류 개선 대상의 계획대비 수행현황을 기술한다. - 오류유형 별 전체 오류대상 수, 최우선 개선대상 수, 단기 및 중장기 개선대상 수 요약
우선 개선과제 정의	<ul style="list-style-type: none"> • 품질개선의 우선 개선 과제에 대해 기술한다. - 우선순위, 개선 수행방식, 개선과제 관련 내용
품질개선 수행 결과	<ul style="list-style-type: none"> • 품질개선 결과에 대한 내용을 기술한다. - 저장영역, DBMS 별 오류발생유형(값, 구조, 표준, 관리체계, 응용프로그램) 별 개선 대상 수/개선 완료 수에 대한 내용

(산출물 : [서식-13]품질개선 결과서)

I

II

III

IV

데이터 품질관리 방법론

V

01

개요



품질통제(Control)는 개선 사업의 효과에 대한 결과를 평가하고, 지속적인 데이터 품질을 유지하고 더 높은 품질을 확보하기 위한 통제활동을 수행하는 단계이다.

품질통제의 세부 절차는 개선결과 평가, 품질목표 관리, 품질통제 실시 단계로 수행한다.

그림 58 품질통제(Control) 절차



표 55 품질통제(Control) 절차 설명

단계	내 용
C1. 개선결과 평가	<ul style="list-style-type: none">기존 품질진단 결과를 개선하여 개선 영역에 대하여 평가한다.- 산출물 : [서식-14] 품질통제 계획서
C2. 품질목표 관리	<ul style="list-style-type: none">개선활동이 지속적으로 이루어질 수 있도록 품질통제 대상을 선정하고 목표를 설정한 후, 품질유지 및 품질확보 방안을 수립한다.
C3. 품질통제 실시	<ul style="list-style-type: none">품질통제 대상에 대해 모니터링 및 관리활동(교육, 홍보 등)을 수행한다.

02

개선결과 평가



개선결과 평가는 목표한 과업범위에 대한 달성률과 개선 전·후를 비교하여 개선결과에 대해 평가한다. 개선 전·후를 비교하여 보여주는 방법은 상당히 직관적인 개선 효과를 보여준다.

▶ 개선결과 비교·평가

개선 영역별, 평가지표별 등으로 개선결과를 비교·평가하여 보여 줄 수 있다.

표 56 개선 영역별 개선 결과(예시)

개선 영역 구분	오류 유형별 비율	
	개선 전	개선 후
관리체계 수립	15.0%	0%
표준 수립	10.0%	0%
구조 보정	15.0%	1.5%
값 보정	25.5%	0.8%

(*비율 : 개선기회 총 건수(100%) 기준)

표 57 평가지표별 달성률 개선 전·후 비교 평가 (예시)

품질지표	평가지표	오류율(전체 건수 기준)	
		개선 전	개선 후
일관성	표준 일관성	1.5%	0%
	참조 무결성	2.0%	0%
	...		
유효성	범위 유효성	1.2%	0%
	날짜 유효성	2.4%	0%
	형식 유효성	1.3%	0%
정확성	논리관계 정확성	0.5%	0.2%
	파생 정확성	0.5%	0.1%
완전성	구조 완전성	0.5%	0%
	...		

(*오류율 : 진단대상 항목별 총 데이터 건수 합 기준 오류건수 합에 대한 품질지표별 비율)

I

II

III

IV

데이터 품질진단 방법론

V

03

품질목표 관리



품질목표 관리는 개선 적용 영역의 품질유지 활동과 개선 미적용 영역에 대한 개선 적용활동이 지속적으로 실행될 수 있도록 체계적인 품질 목표를 수립하여 관리하고, 달성 현황은 정량적으로 확인한다.

▶ 품질목표 달성 현황 정량적 확인

품질관리 개선 영역별 품질목표를 달성해 나가는 현황을 정량적으로 확인되어야 한다.

표 58 품질관리 개선 영역별 품질목표 달성 정량적 확인 사항

관리 영역	세부 사항
품질관리체계 개선	<ul style="list-style-type: none"> 데이터 품질 관리수준 지표별 체크 리스트를 활용하여 품질 관리체계 수준의 변화하는 모습을 주기적으로 확인
표준화 개선	<ul style="list-style-type: none"> 전체 대상 중 표준 적용 비율을 확인하여 표준이 적용되어 나가는 모습을 주기적으로 확인
값 보정	<ul style="list-style-type: none"> 데이터 값의 오류율과 실제 보정이 적용되는 영역의 비율을 확인하여 적용되어 나가는 모습을 주기적으로 확인
구조 보정	<ul style="list-style-type: none"> 구조의 오류율과 실제 보정이 적용되는 영역의 비율을 확인하여 적용되어 나가는 모습을 주기적으로 확인
응용프로그램 보정	<ul style="list-style-type: none"> 응용프로그램의 오류율과 실제 보정이 적용되는 영역의 비율을 확인하여 적용되어 나가는 모습을 주기적으로 확인

개선 완료된 영역에 대해서는 데이터 품질유지 방안을 수립하여 품질수준이 유지될 수 있도록 품질이슈에 대한 모니터링 등 지속적인 사전예방활동이 강화되어야 한다.

▶ 품질통제 계획서 작성

지속적인 품질관리 활동을 위해 개선결과 품질수준 평가와 품질수준 목표설정 등이 구성된 품질통제 계획서를 작성한다.

표 59 품질통제 계획서 구성

항 목	내 용
품질통제 목적	<ul style="list-style-type: none"> 품질통제 수행의 목적을 기술한다. <p>예) 시스템 품질진단 정기 점검 결과 오류에 대한 품질 개선 효과를 평가하고, 지속적인 데이터 품질을 유지하고 더 높은 품질을 확보하기 위한 통제 활동 수행</p>
품질수준 평가	<ul style="list-style-type: none"> 개선 전후의 오류율 및 성숙도 평가를 통해 현 수준의 데이터 품질수준을 평가한다. <p>1) 지표별 오류율 개선 결과 - 평가지표별 오류율의 개선 전·후 내용</p> <p>2) 지표별 성숙도 평가 내용</p>
품질수준 목표설정	<ul style="list-style-type: none"> 품질 수준 목표에 대해 기술한다. <p>- 평가지표별 품질목표(오류율 및 성숙도) 내용</p>
품질통제 활동 계획	<ul style="list-style-type: none"> 품질 수준 목표 달성을 위한 품질활동 및 모니터링 계획에 대해 기술한다. <p>예) 지속적인 품질개선 및 모니터링 활동, 품질관리 교육 및 문화 정착, 품질오류 신고 접수 및 처리 활동 등</p>

(산출물 : [서식-14]품질통제 계획서)

I

II

III

IV

데이터 품질진단 방법론

V

04

품질통제 실시



품질통제는 품질관리가 지속적으로 수행될 수 있도록 하는 다양한 활동을 의미한다.

표 60 품질통제 활동

통제 활동 구분	상세 내용
지속적인 품질개선 및 모니터링 활동	<ul style="list-style-type: none"> 개선 가이드 내용에 대해서 실제 개선이 되었는지 지속적으로 확인하여 현재 품질 수준이 목표 대비 어느 정도인지 주기적으로 확인하고, 실제 개선에 대한 독려 활동
인식 전환을 위한 교육 활동	<ul style="list-style-type: none"> 조직 구성원들의 품질에 대한 인식전환을 위해 체계적인 교육이 필요하며, 일회성이 아닌 지속적인 교육 활동 필요
공감대 형성 및 문화정착을 위한 활동	<ul style="list-style-type: none"> 데이터 품질의 중요성에 대한 공감대를 형성하고 절차를 철저히 수행하는 문화를 정착시키기 위한 다양한 변화관리 활동
품질오류 신고 접수 및 처리에 관한 활동	<ul style="list-style-type: none"> 수집된 품질이슈 내용을 확인하고 즉각 처리하도록 하며, 처리가 지연될 시에는 이에 대해 신고자에게 통지하고 조속히 처리될 수 있도록 방안을 강구 - 중대한 품질오류신고에 대해서는 필요 시, 태스크포스(Task Force)와 같은 임시조직을 만들어 대응할 수도 있다.

▶ 데이터 품질을 유지하는 가장 효율적인 방법

표 61 데이터 품질을 유지하는 가장 효율적인 방법

전사 구성원	① 품질관리 절차에 따라 업무를 수행하도록 한다.
업무 담당자 (데이터 담당자)	② 품질이슈에 대한 관심을 가지도록 수시로 피드백을 주어 대응하도록 한다.
정보화 담당자 (데이터 처리자)	③ 시스템 구축 및 도입 시, 철저한 품질관리 활동을 수행하도록 하여 ④ 동일한 품질이슈가 다시 발생되지 않도록 한다.

빅데이터 플랫폼 및 센터 데이터 품질관리 가이드

2020년 2월 초판 인쇄

2020년 2월 초판 발행

발행인 문 용 식

발행처 한국정보화진흥원

대구광역시 동구 첨단로 53

본 가이드는 2019년 한국정보화진흥원의 「빅데이터 플랫폼 및 센터 데이터품질관리」 용역 사업의 결과 산출물입니다.

본 가이드 내용의 무단 전재 및 재배포를 금하며, 인용 시에는 반드시 자료의 출처인 한국정보화진흥원 “빅데이터 플랫폼 및 센터 데이터 품질관리 가이드” 라고 밝혀주시기 바랍니다.

본 가이드의 내용과 관련한 문의는 한국정보화진흥원 지능데이터본부(지능데이터기반팀)로 연락주시기 바랍니다.

