# 데이터 품질진단 절차 및 기법 (Ver 1.0)

## Data Quality Assessment Procedure Manual

데이터 품질진단을 위한 실무 지침

정형 데이터와 비정형 데이터의 품질 개선을 위한
품질 측정 및 분석 기법

KDB Korea Database Agency 한국데이터베이스진흥원

# 머리말

정보통신 기술의 발달은 기업의 경영 방식에 커다란 변화를 일으켰다. 글로벌 경쟁력을 갖춘 기업은 연구개발, 제조 및 영업 부문간에 데이터를 장벽 없이 실시간으로 공유하며 각 부문의 데이터를 기반으로 경영층의 신속하고 정확한 의사결정을 지원한다. 현대 기업은 정보통신 기술에 기반을 둔 객관적인 데이터에 의존하지 않으면 존립하기 곤란하다. 부정확하고 시의 적절치 못한 데이터는 비합리적이고 경쟁사보다 늦은 의사결정을 유도할 것이기 때문이다.

지금까지 국가나 기업의 정보화 계획은 네트워크, 하드웨어 및 소프트웨어를 중심으로 이루어져 온 것이 사실이다. 대체로 의사결정을 하는 경영층은 데이터가 일단 컴퓨터에 입력되면, 컴퓨터에 입력된 데이터라는 이유만으로 매우 정확할 것이라는 막연한 인식을 갖고 있다. 그에 따라 자연스럽게 데이터에 대한 관심이 상대적으로 위축되었고 정보화 초기에는 데이터의 규모도 크지 않아 데이터 품질은 국가나 기업의 정보화 정책에서 외면당해 왔다.

그러나 최근 인터넷을 통해 유통되는 가비지 데이터(Garbage Data)에 대한 피해 사례나 잘못된 데이터로 인한 영향이 얼마나 심각한지는 신문이나 방송을 통해 많이 알려지고 있다. 특히 최근에 금융 데이터나 국가의 행정 데이터에 생긴 오류로 인해 관련된 개인이나 조직에 상당한 피해를 주었다는 소식이 빈번하다. 데이터가 정보 시스템의 부속품으로 여겨지는 시대는 지났다.

이제 데이터 품질관리는 정보 시스템에서 필수 불가결한 기반 프로세스이다. 특히 데이터 품질진단은 데이터 품질관리를 위한 첫 걸음이다. 자사의 정보 시스템이 운영하는 데이터 중 어느 부분에 오류가 있고, 그 원인이 무엇인지를 파악하는 것이 품질관리 활동의 가장 기초적인 업무이다. 그 이유는 데이터의 품질관리를 위한 프로세스와 조직의 개선 방향은 현재 데이터의 현상에 따라 큰 차이가 있을 수 있기 때문이다.

본 지침서에서는 데이터 품질진단 절차의 기초적인 개념을 가급적 단순화하여 정리하였고, 데이터 품질진단 기법을 정형 데이터와 비정형 데이터로 구분하여 수록함으로써 최근 증가 추세에 있는 콘텐츠 데이터베이스의 품질진단에 적용할 수 있도록 노력하였다. 본 지침서가 국내 데이터 품질관리 담당자들에게 도움이 되기를 기대한다.

# 목 차

# 그림 목차

## 제2장 데이터 품질진단 절차



## 제3장 데이터 품질진단 기법 – 정형 데이터

## 제4장 데이터 품질진단 기법 – 비정형 데이터

# 표 목차

## 제2장 데이터 품질진단 절차



## 제3장 데이터 품질진단 기법 – 정형 데이터

## 제4장 데이터 품질진단 기법 – 비정형 데이터

# 제1장 데이터 품질진단 개요

## 제1절 데이터 품질진단 지침의 목적

데이터는 일상 업무에 있어서 이미 절대적으로 필요한 존재가 되었으며 업무를 수행하는 과정에서 수많은 데이터가 생산되기도 하여 이러한 데이터를 어떻게 분류하여 체계적으로 보관하고 활용하느냐 하는 문제는 정보시스템의 최대 이슈이기도 하다. 2008년 IDC 보고서에 따르면 한국의 디지털 정보량은 연평균 57% 정도 증가하는 것으로 분석되었으며, 2011년에 이르면 디지털 정보량은 대략 2만7천 페타바이트(PB, 1PB는 약 100만 기가바이트)에 달할 것으로 전망하고 있다.

한편 디지털 정보는 이미지, 음성 등 비정형 데이터와 일반적인 정형 텍스트 데이터로 구분해 볼 수 있는데, 이 중 비정형 데이터가 전체 정보량의 92%를 차지하고 있고, 정형 텍스트 데이터는 8%에 불과한 것으로 조사되어 비정형 데이터가 디지털 정보에서 압도적인 비중을 차지하고 있음을 알 수 있다. 그러나 현재까지 데이터에 대한 품질관리는 정형화된 텍스트 데이터에 편중되어 있고, 동영상, 이미지, 3D 등 비정

형·멀티미디어 콘텐츠에 대해서는 매우 취약하며, 비정형 데이터와 정형 데이터에 대한 품질을 일관성 있게 통합적으로 관리할 수 있는 방법이 부재하다.

이에 본 지침은 정형 데이터와 비정형 데이터를 아우르는 통합적인 품질관리 방법을 개발·제안하여, 정보시스템 전체 데이터의 품질 향상을 도모하고 나아가 조직의 업무 효율성 제고 및 고객의 콘텐츠 활용 만족도 향상에 기여함을 목적으로 한다.

## 제2절 데이터 품질진단의 정의와 종류

데이터 품질진단은 해당 조직이 운영·관리하고 있는 정보시스템에 저장된 정형·비정형 데이터의 품질을 측정하여 현재의 수준을 평가하고 품질 저하의 요인을 분석하여 개선 사항을 제안하는 절차이다. 따라서 데이터 품질진단은 데이터 품질관리 활동의 일환으로 전개되는 것이 바람직하다.

데이터와 관련된 품질진단의 종류는 크게 데이터 값 진단, 데이터 구조 진단, 데이터 관리 프로세스 진단 등으로 구분할 수 있다.

정형 데이터에 대한 데이터 값 진단은 운영 데이터베이스의 테이블·컬럼·코드·관계·업무규칙을 기준으로 데이터의 값에 대한 현상을 분석한다. 또한 데이터 값과 관련된 품질기준을 적용하여, 오류내역을 산출하고 주요 원인을 분석하여 개선 사항을 제안한다. 특히 데이터 값과 관련된 오류는 데이터의 구조·흐름 통제·관리 프로세스와 연관되어 발생한다. 초기 구조 설계 오류·일관성 결여·관리 프로세스 결여·오너쉽 결여 등으로 인하여 낮은 품질의 데이터가 지속

적으로 발생할 수 있다. 따라서 데이터 값 진단의 개선사항은 그 오류발생 원인 분석에 따라 값의 정제 외에도 구조 개선사항・데이터 흐름통제・관리 프로세스의 개선사항이 포함된다.

비정형 데이터에 대한 데이터 값 진단은 비정형 콘텐츠 자체의 상태와 메타데이터에 대한 데이터 품질진단으로 이루어진다. 비정형 콘텐츠 자체의 상태에 대한 진단은 그 내용 자체의 합목적성을 비롯하여 동영상, 이미지, 3D 등 비정형 콘텐츠 유형 따라 각기 다른 관점에서 작성된 콘텐츠의 상태를 시각이나 청각, 또는 자동화된 도구를 이용하여 진단한다. 메타데이터의 경우는 콘텐츠 파일 자체에 저장되는 메타데이터와 정형 데이터의 형식으로 데이터베이스에 별도 저장되는 메타데이터로 구분하여 주로 데이터베이스에 별도 저장된 메타데이터에 대하여 정형 데이터와 유사한 기준과 방법에 의해 진단을 수행한다. 또한 정형 데이터에 대한 진단에서와 마찬가지로 비정형 콘텐츠 유형별로 관련된 품질기준을 적용하여, 오류내역을 산출하고 주요 원인을 분석하여 개선 사항을 제안하는 절차가 포함된다.

데이터 구조 진단은 데이터 모델링 관점에서 데이터 품질을 진단한다. 특히 중요 업무 데이터베이스의 리버스 모델링(Reverse Modelling)을 통하여 논리 모델을 작성하고, 현행 데이터베이스의 구조 무결성・데이터 구조 표준화・관리 수준・변경 관리 등의 현황을 진단한다. 구조 진단의 주요 이슈는 데이터의 표준화 수준, 표준 코드, 표준 도메인, 테이블·컬럼 및 관계 정의, 정규화 수준 등이 이에 해당된다. 데이터 구조 진단은 주로 정형 데이터에 대해 이루어지나 비정형 데이터의 메타데이터에 대해서도 수행될 수 있다.

데이터 관리 프로세스 진단은 정형・비정형 데이터에 대한

현행 데이터 관리 프로세스를 분석하여 문제점을 도출하고, 이를 개선할 수 있는 핵심 업무 프로세스를 표준화하여 재설계한다. 이때 조직 역량분석을 통해 단계적 품질관리 이행 전략을 수립하는 절차를 제안할 수 있다. 데이터 관리 프로세스 진단과 관련된 주요 이슈는 품질관리 정책 수립·업무 프로세스의 적절성 및 운영성 분석·프로세스별 오너쉽 등으로 볼 수 있다.

데이터 값, 데이터 구조 및 데이터 관리 프로세스가 모두 건실해야 데이터 품질을 보장할 수 있다. 데이터 구조나 관리 프로세스에 문제가 있으면 당연히 데이터 값의 품질이 떨어지고, 데이터 관리 프로세스가 없는 조직에서 높은 품질의 데이터 값이나 구조를 기대할 수 없다. 데이터 값, 데이터 구조 및 데이터 관리 프로세스는 데이터 품질 측면에서 상호 긴밀한 관계를 갖는다.

## 제3절 지침의 범위와 구성

본 지침에서 다루고자 하는 품질진단 대상은 좁은 의미의 데이터이다. 즉, 조직이 운영하는 정보 시스템에 구축된 정형·비정형의 디지털 데이터이다. 또한 데이터의 품질진단의 종류도 전술한 데이터 값, 데이터 구조 및 데이터 프로세스 관리 중 데이터 값에 한정하여 다루고자 한다. 또한 품질진단 대상이 되는 데이터 값은 데이터의 외형적 형태에 관심을 두고 있으며, 데이터 값의 내재적인 의미는 보고서 범위의 밖이다. 물론 정형·비정형 데이터의 의미론적 품질진단이나 평가도 매우 중요한 과제이나 본 지침서에서는 정형 데이터의 경우 가급적 데이터 품질진단 결과를 데이터 품질진단 전용 소프트웨어나 SQL에 의거하여 객관적이고 정량적으로 도

출하는 것에 중심을 두고 있으며, 비정형 데이터의 경우는 내용 자체의 질적 우수성보다 구축 목적에 대한 부합성과 함께 시각이나 청각, 또는 전용 소프트웨어나 SQL을 이용하여 콘텐츠의 작성 상태와 메타데이터의 상태를 객관적·정량적으로 도출하고자 하는데 무게를 두고 있다.

지침은 크게 세 부분으로 구성되어 있다. 제1장에서는 데이터 품질진단 지침 개발 목적과 주요 개념에 대한 정의와 종류를 소개하였다. 제2장에서는 데이터 품질진단의 절차에 대하여 설명하였다. 데이터 품질진단 절차로서 크게 계획 수립·기준 및 진단 대상 정의·품질측정 및 분석·품질 개선 활동 등으로 구분하였다. 본 장에서는 절차별로 2003년부터 한국데이터베이스진흥원의 데이터 품질진단 실무진들의 경험과 기존의 관련 문헌, 관련 연구 수행 결과를 바탕으로 하여 핵심 절차를 중심으로 소개하였다. 제 3장과 제 4장에서는 데이터 품질진단 절차에 대한 이해를 높이기 위한 관련 기술이나 방법론, 사례 등을 소개하였다. 정형 데이터에 대한 진단의 경우 실무에 적용하는데 도움을 주고자 필요시 관련 주제에 대한 SQL 문을 함께 제시하였으며, 비정형 데이터에 대한 진단의 경우에는 실무에 도움이 될 수 있도록 체크리스트 예문을 함께 제시하였다.

# 제2장 데이터 품질진단 절차

본 장에서는 데이터 품질진단 절차를 품질진단 계획 수립, 품질기준 및 진단 대상 정의, 품질측정, 품질측정 결과 분석, 데이터 품질 개선 등의 5단계로 구분하였다. 해당 진단 절차 및 세부 수행 활동은 진단을 수행할 데이터의 정형·비정형 특성에 따라 적용할 수 있는 절차와 적용하기 어려운 절차가 존재하므로, 데이터의 형태별 특성에 따라 조정하여 적용해야 한다.

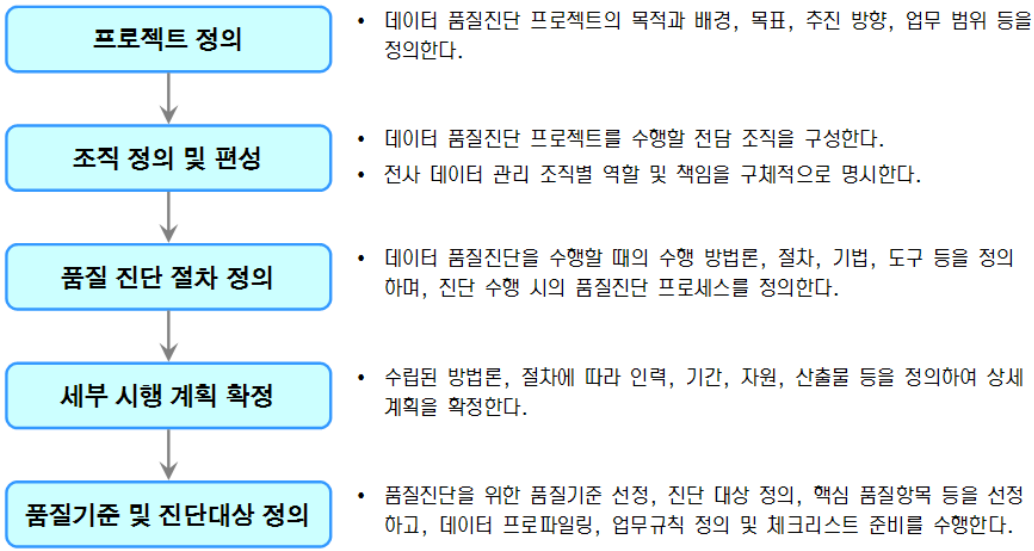
## 제1절 품질진단 계획 수립

### 1. 데이터 품질진단 프로젝트 정의

품질진단 계획수립은 크게 두 개의 프로세스로 구분해 볼 수 있는데, 그 하나는 데이터 품질진단을 본격적으로 수행하기에 앞서 사전에 진행할 품질진단의 개요 및 정의, 데이터 품질과 관련 이용자의 요구사항 파악, 품질진단 수행 조직의 정의, 세부 수행절차 확립 등의 계획을 수립하는 과정이고, 다른 하나는 실제로 품질측정을 수행하기 위한 품질기준 및

대상 선정, 데이터 프로파일링, 업무규칙 정의 및 체크리스트 준비 등이 이루어지는 과정이다.

품질진단 계획 수립은 데이터 품질진단을 수행하기 위한 사전 준비 절차로서 현재 운영하는 정보 시스템에 대하여 조직 내부의 담당자들이 인식하고 있는 정형·비정형 데이터의 품질 문제를 사전에 파악하여 조직의 품질관리 목표와 범위를 정의하고, 소요 비용·인력·조직·보유 S/W·장비 등을 분석하여 가용 자원을 확보하며, 이를 바탕으로 세부 수행계획을 정립해야 한다. 전반적인 절차는 〈그림 2-1〉과 같다.




〈그림 2-1〉 품질진단 계획수립 절차

## 1.1 개요

데이터 품질진단 프로젝트가 효과적으로 소정의 목적을 명확하게 달성할 수 있도록 프로젝트의 배경, 목표, 추진방향, 범

위 등을 정의한다.

## 1.2 수행절차

1) 업무 담당자와의 면담을 통하여 품질진단 프로젝트의 목적을 설정한다. 목적 설정시 다음 사항을 고려한다.
   - 데이터 관리 조직
   - 데이터 품질관리 목표 및 배경
   - 데이터 품질 요구사항
   - 조직의 품질진단 결과 활용 계획
   - 향후 데이터 품질 개선 계획
2) 설정된 목적에 따라 관리 대상 시스템, 업무, 정형·비정형 데이터 현황을 파악하여 진단을 수행할 대상을 선별하고 이를 진단 범위로 확정한다.
3) 업무담당자 및 IT부서 담당자와 면담을 통해 조직의 품질진단 여건을 사전 파악하고 품질진단 프로젝트에 소요될 자원을 파악한다.
   - 데이터 품질진단 프로젝트에 책정된 예산 규모
   - 지원 가능한 인력 규모
   - 품질진단 대상 시스템의 운영 데이터의 규모 및 특성
   - 품질진단 도구 및 진단 기법
   - 품질진단 수행 인력 구성 계획
   - 품질진단 수행 기간 : 수행 가능한 시작일 및 종료일

## 1.3 산출물 : 데이터 품질진단 프로젝트 정의서

### 1.4 참고사항

데이터 품질진단 프로젝트의 목적과 범위는 프로젝트 수행 후 기대효과를 만족시킬 수 있도록 고려되어야 한다. 품질진단 프로젝트를 수행하는 조직의 목적에 따라 수반되는 작업 유형과 프로젝트의 범위가 결정된다. 예를 들어 품질진단을 수행하는 주된 목적이 품질관리 개선 이전 단계의 선행적 데이터 결함 수준 파악일 경우에는 데이터 프로파일링을 적용한 데이터 현상 분석으로 프로젝트의 범위가 제한될 수도 있다.

## 2. 수행 조직 정의

### 2.1 개요

현행 데이터 관리 조직을 분석하여, 데이터 품질진단 프로젝트를 수행할 전담 조직을 정의한다. 전담 조직은 조직의 데이터 품질관리 원칙과 개선 목표에 부합되고 프로젝트의 규모에 적합해야 하며, 〈그림 2-2〉와 〈표 2-1〉을 참조하여 정의한다.

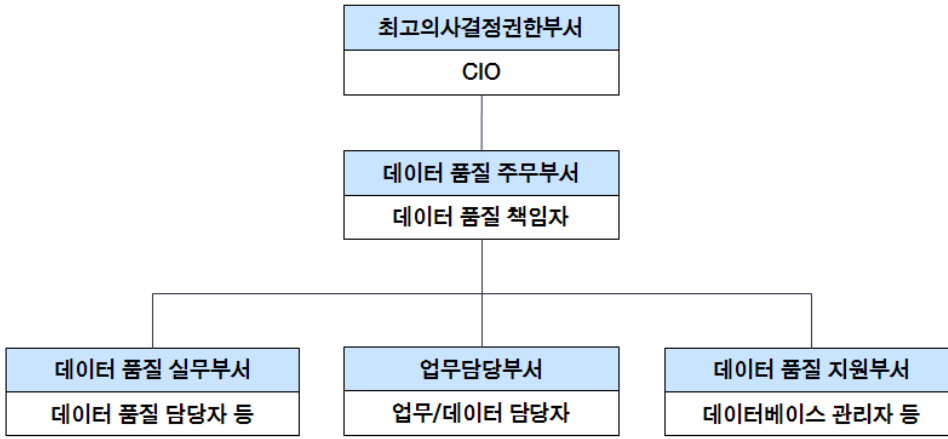
### 2.2 수행절차

1) 현행 품질진단을 수행할 데이터 관리 조직의 규모, 인력, 역할을 분석한다.
2) 데이터 품질진단을 효율적으로 수행하기 위한 품질 전담 조직을 검토하고, 현 데이터 운영 조직의 여건과 규모에 부합된 진단 수행 조직을 정의한다.
3) 해당 조직을 정의한 후, 해당 업무별 역할과 책임한계를 구체적으로 정의한다.

### 2.3 산출물 : 데이터 품질진단 조직 구성도

### 2.4 참고사항

데이터의 품질진단과 이에 따른 개선활동을 효과적으로 수행하기 위해서는 체계적인 조직 구성이 필요하다. 데이터 품질진단 및 개선은 전사적 관점에서 계획되고 수행되어야 하기 때문에 〈그림 2-2〉에서 보는 바와 같이 최고 의사결정 권한자로부터 해당 업무 또는 데이터에 대한 담당자에 이르기까지 체계적인 조직 구성과 이에 따른 업무 수행이 요구된다.




〈그림 2-2〉 데이터 품질진단 수행 조직 구성

〈표 2-1〉 데이터 품질진단 수행조직의 역할과 책임

| 구 분 | 역 할 | 책 임 |
|---|---|---|
| 최고의사 결정권한부서 | 정형·비정형 데이터에 대한 품질진단 및 개선과 관련한 최고 의사결정권한 수행 | • 데이터 품질정책 결정<br>• 품질관리 우선순위 및 목표 설정<br>• 품질개선을 위한 의사결정 |
| 데이터 품질 주무부서 | 정형·비정형 데이터의 품질진단과 관련하여 최고의사결정권한자의 결정사항을 집행하고 제반 정책적 결정이 필요한 사안들을 취합하며, 데이터 품질진단 전반에 대한 실무적 측면의 책임과 의사결정을 수행 | • 데이터 품질정책에 따른 구체적인 품질목표 및 추진방향 설정<br>• 품질진단 계획 승인 및 진단결과 확인<br>• 품질개선 계획 승인 및 개선결과 확인 |
| 데이터 품질 실무부서 | 품질진단에 대한 실무를 전담하며, 품질진단을 위한 품질기준과 데이터 프로파일링, 업무규칙 도출, 품질측정, 측정결과분석 및 개선활동 관리 등을 수행 | • 품질기준 및 품질지수 관리<br>• 품질측정 항목 관리<br>• 품질측정 대상 선정<br>• 데이터 프로파일링 및 업무규칙 도출<br>• 품질측정 수행<br>• 측정결과 및 개선결과에 대한 보고<br>• 개선활동 관리 및 담당자 교육 |
| 데이터 품질 지원부서 | 각 업무/데이터가 구축 목적에 맞게 정상적으로 사용될 수 있도록 대상 개체 및 시스템 인프라를 관리하고 데이터 품질 실무 담당자에 대한 지원을 수행 | • 품질측정 수행 지원<br>• 품질측정 결과분석 지원<br>• 품질개선활동 지원 |
| 업무 담당 부서 | 각 업무별 정형·비정형 데이터에 대해 가장 잘 이해하고 있으면서 해당 데이터의 생산/운영 과정에서 1차적인 품질관리 책임을 수행하고, 품질담당자와 협조 하에 해당 데이터에 대한 품질개선에 주도적인 역할을 수행 | • 소관 데이터에 대한 품질이슈 관리<br>• 소관 데이터에 대한 업무규칙 도출<br>• 업무규칙 준수 및 점검<br>• 소관 데이터와 관련한 품질개선 수행 |

## 3. 품질진단 절차 정의

### 3.1 개요

데이터 품질진단 프로젝트를 수행하기 전에 조직의 현황과 대상 시스템의 규모와 특성에 맞는 데이터 품질진단 절차를 정의한다. 동일한 품질진단 프로젝트를 수행할지라도 적용하는 품질관리 및 진단 절차는 수행 조직의 규모와 성격에 따라 서로 상이할 수 있다. 따라서 조직의 데이터 관리 수준, 인력, 자원 등을 종합적으로 고려하여 현실에 부합된 진단 절차를 정의하고 적합한 품질 방법론이나 프레임워크를 적용하여 적합한 진단 절차를 도출한다.

### 3.2 수행절차

1) 과거 품질진단 사례, 유사 품질진단 프로젝트 사례를 중심으로 품질진단 절차, 방법론, 프레임워크 등 관련 자료를 분석한다.
2) 현 조직의 규모, 인력, 자원에 부합된 품질진단 절차 및 방법론을 선정한다.
3) 선정된 품질진단 절차 및 방법론을 토대로 현조직의 여건에 부합된 기능, 조직, 프로세스를 설계한다.
4) 적용할 품질진단 절차에 대한 적정성을 검토하여 확정한다.

### 3.3 산출물 : 품질진단 절차도

## 4. 세부 시행 계획 확정

### 4.1 개요

사전에 수립된 데이터 품질진단 프로젝트 계획, 조직, 절차를 토대로 품질진단 프로젝트를 성공적으로 수행하기 위한 적합한 인력, 자원, 세부 수행 일정, 예산 등의 배정 및 확정을 통해 프로젝트 계획을 완성한다.

### 4.2 수행절차

1) 사전 정의된 프로젝트의 범위, 요구사항, 조직, 방법론, 수행 절차에 따라 조직 내에서 수행 가능한 일정, 절차, 자원 등을 상세히 정의하여 세부시행 계획서를 작성한다. 세부 시행 계획서에는 다음 사항을 포함하도록 한다.
   - 수행 가능한 프로젝트 일정
   - 세부 추진 절차
   - 전담 조직 구성도
   - 자원배분 계획
   - 이용자 참여계획
   - 교육 및 면담계획
2) 작성된 세부 시행 계획서는 의사 결정권자의 승인을 통하여 공식 문서화하며, 공식화된 일정에 따라 품질 프로젝트를 체계적으로 수행한다.

### 4.3 산출물 : 프로젝트 세부 수행 계획서

# 제2절 품질기준 및 진단 대상 정의

품질기준 및 진단 대상 정의는 실제로 품질측정을 수행하기 위한 품질기준 및 대상 선정, 데이터 프로파일링, 업무규칙 정의 및 체크리스트 준비 등이 이루어지는 과정으로, 데이터 품질진단 수행 시 적용할 품질측정의 기준을 사전에 정의하고, 데이터의 형태(정형·비정형)에 따라 데이터 품질진단을 수행할 대상 정보 시스템의 테이블 및 컬럼 등을 정의하거나 진단 대상이 되는 멀티미디어 콘텐츠 및 해당 메타데이터를 선정하여, 데이터 유형별 특성에 따른 데이터 프로파일링 및 업무규칙 정의, 체크리스트 준비 등을 수행한다.


〈그림 2-3〉 데이터 품질기준 및 진단대상 정의 절차

## 1. 데이터 품질기준 선정

### 1.1 개요

데이터 품질기준은 데이터의 품질 수준을 평가하는 기준으로 데이터의 정확성 확보를 위하여 지속적으로 관리되어야 할 측정 기준을 의미한다. 데이터 품질기준은 객관성 확보가 중요하므로 데이터 품질과 관련된 이해당사자 및 조직원이 공동 협의하여 도출한다.

### 1.2 수행절차

1) 선행 품질기준 관련 자료를 조사한다. 특히 품질 평가기준 연구 혹은 유사 프로젝트를 조사하여 조직 내부에서 적용 가능한 후보 품질기준을 수집한다.
2) 조직 내에서 활용할 품질기준을 실제 품질측정 절차를 고려하여 선정하고, 세부 품질기준을 도출한다.
3) 업무담당자, 품질팀, IT부서 담당자 등의 공동 협의를 거쳐 최종 확정한다.

### 1.3 산출물 : 데이터 품질기준 정의서

### 1.4 참고사항

데이터의 품질기준은 품질의 기대수준을 요구하는 조직의 관점・이용자의 관점・이해 관계자의 관점에 따라 각각 그 중요도가 상이하고 그 해석이 다양하다. 따라서 과거 데이터 품질과 관련된 품질 연구 결과나 외부 품질 전문가들의 견해를 수렴하여 조직의 경영목표와 품질관리 정책에 부합된 품질기준을 선별하여 활용하는 것이 바람직하다. 품질기준은

그 용어와 의미가 명확하게 정의되어야 하며 누구나 이해하기 쉽게 적용 사례를 활용하여 표현되어야 한다.

### 1) 정형 텍스트 데이터에 대한 일반적인 데이터 품질기준

국내외 데이터 품질 관련 문헌이나 데이터 품질관리 컨설팅 기관의 자료를 종합적으로 정리하면 정형 텍스트 데이터에 대한 완전성・유일성・유효성・일관성・정확성의 5개의 품질기준에 대해 〈표 2-2〉와 같이 정의할 수 있다.

〈표 2-2〉 일반적 데이터 품질기준 정의

| 품질기준 | 정 의 |
|---|---|
| 완전성 (Completeness) | 필수항목에 누락이 없어야 한다. |
| 유일성 (Uniqueness) | 데이터 항목은 유일해야 하며 중복되어서는 안 된다. |
| 유효성 (Validity) | 데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 한다. |
| 일관성 (Consistency) | 데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의되고, 서로 일치해야 한다. |
| 정확성 (Accuracy) | 실세계에 존재하는 객체의 표현 값이 정확히 반영이 되어야 한다는 것을 의미한다. |

완전성・유일성・유효성・일관성・정확성 등 5개의 일반적인 데이터 품질기준이 조직에서 필요한 품질관리 실무 기준을 활용하기에 부족할 수 있다. 이를 해결하기 위하여 도출된 데이터 품질기준을 보다 상세화하고 그 의미를 세분화하여 하위 품질기준을 정의할 수 있다. 〈표 2-3〉은 5개의 일반적인 데이터 품질기준을 대상으로 그 적용기준을 확장하여 상

〈표 2-3〉 세부 품질기준 설명 및 활용 예시

| 품질 기준 | 세부 품질기준 | 품질기준 설명 | 활용 예시 |
|---|---|---|---|
| 완전성 | 개별 완전성 | 필수 컬럼에는 값의 누락이 없어야 한다. | 고객의 아이디는 NULL일 수 없다. |
| | 조건 완전성 | 조건에 따라 컬럼 값이 항상 존재해야 한다. | 기업 고객의 사업자 등록번호가 NULL일 수 없다. |
| 유효성 | 범위 유효성 | 컬럼 값이 주어진 범위내에 존재해야 한다. | 기준점 좌표각은 -360 초과 360미만까지의 값을 가진다. |
| | 날짜 유효성 | 컬럼 값이 날짜유형일 경우에는 유효날짜 값을 가져야 한다. | 99991231, 20080231은 유효하지 않은 값이다. |
| | 형식 유효성 | 컬럼은 정해진 형식과 일치하는 값을 가져야 한다. | 주민번호형식은 '999999-9999999'의 형식이어야 한다. |
| 정확성 | 선후관계 정확성 | 복수의 컬럼값이 선후 관계에 있을 경우 이 규칙을 지켜야 한다. | 시작일은 종료일 이전 시점이어야 한다. |
| | 계산/집계 정확성 | 한 컬럼의 값은 다수 컬럼의 계산된 값일 경우 계산 값이 정확해야 한다. | 월 통계 테이블의 매출액은 기산일이 현월인 매출액의 총합과 일치해야 한다. |
| | 최신성 | 정보의 발생, 수집, 그리고 갱신 주기를 유지해야 한다. | 고객의 현재 값은 고객변경이력의 마지막 ROW와 일치해야 한다. |
| | 업무규칙 정확성 | 컬럼이 업무적으로 복잡하게 연관된 경우 관련 업무규칙에 일치해야 한다. | 지급원장의 지급여부가 'Y'이면 지급원장의 지급일자는 신청일보다 이전시점 이어야 하고 NULL이 아니어야 한다. |
| 유일성 | 단독 유일성 | 컬럼은 유일한 값을 가져야 한다. | 고객의 이메일 주소는 유일해야 한다. |
| | 조건 유일성 | 업무 조건에 따라 컬럼 값은 유일해야 한다. | 교육과정의 강의시작일이 있으면, 강의실 코드, 임대일, 강사코드가 모두 동일한 레코드는 존재하지 않는다. |
| 일관성 | 기준코드 일관성 | 컬럼이 통합코드를 기준코드로 사용할 경우 그 참조무결성을 유지해야 한다. | 고객의 직업코드는 통합코드테이블의 직업코드에 등록된 값이어야 한다. |
| | 참조무결성 | 데이블간의 컬럼값이 참조관계에 있는 경우 그 무결성을 유지해야 한다. | 대출원장의 대출원장번호는 대출 상세내역에 존재해야 한다. |
| | 데이터 흐름 일관성 | 데이터를 생성하거나 가공하여 데이터가 이동되는 경우, 연관된 데이터는 모두 일치해야 한다. | 운영계의 현재 가입 고객 수와 DW의 고객 수는 일치해야 한다. |
| | 컬럼 일관성 | 관리 목적으로 중복 컬럼을 임의 생성하여 활용하는 경우 각각의 동의어 컬럼 값은 일치해야 한다. | 주문의 주문번호와 고객번호는 배송의 주문번호와 고객번호가 서로 일치해야 한다. |

세 품질기준을 도출한 사례이다. 상세 품질기준을 도출하는 데 있어서 중요한 것은 도출된 기준이 실제 측정 가능해야 한다는 것이다. 따라서 품질기준의 측정 방법과 결과물들을 사전에 조사하여 명확히 측정 가능한 품질기준만을 선정해야 한다.

### 2) 비정형 데이터에 대한 데이터 품질기준

비정형 데이터에 대한 품질기준은 정형 텍스트 데이터에 대한 품질기준과 다소 다르게 적용되어야 한다. 비정형 데이터는 디지털화된 멀티미디어 콘텐츠를 지칭하는 것으로, 이를 좀 더 세분하면 비정형 콘텐츠 자체와 그 메타데이터로 나누어 볼 수 있다. 비정형 콘텐츠의 메타데이터는 정형 텍스트 데이터와 품질 특성이 거의 같으나 일반적인 정형 텍스트 데이터가 현실 세계에 존재하는 개체나 사건에 대한 내용을 텍스트 데이터로 표현하고 있는데 반해 비정형 콘텐츠의 메타데이터는 시스템 내부에 디지털화 되어 존재하는 멀티미디어 개체를 표현하고 있다는 점에서 차이가 있다. 이러한 특성 차이 때문에 비정형 콘텐츠의 메타데이터는 그 데이터가 대상 개체를 정확하게 식별해 낼 수 있는지, 즉 데이터와 실물 콘텐츠 개체 간의 일치성으로까지 무결성의 개념이 확장되어야 하고, 콘텐츠 자체에 대한 변경이 해당 메타데이터에 적절하게 반영되고 있는지도 최신성 또는 적시성 관점에서 확인되어야 한다. 비정형 콘텐츠의 메타데이터에 대한 품질기준은 〈표 2-3〉에서 설명한 품질기준에 대해 이러한 차이점을 추가로 고려해야 한다.

비정형 콘텐츠 자체에 대한 품질기준은 콘텐츠 유형에 따라 다소 다를 수 있다. 일반적으로 비정형 콘텐츠의 유형을 구분하는 기준은 여러 학자나 연구내용, 활용 목적 등에 따라

다를 수 있으며, 〈표 2-4〉은 데이터베이스 구축 관점에서 콘텐츠 유형을 분류한 사례이다. [1)]

〈표 2-4〉 콘텐츠 유형 분류 사례

| 대상 자료 유형 | | 내 용 |
|---|---|---|
| 메타데이터 | | 콘텐츠에 대한 각종 정보를 가지고 있는 데이터로 구축되는 DB 형태 |
| 텍스트 | 직접입력 방식 | 문자의 직접 입력 작업으로 구축되는 DB 형태 |
| | OCR 변환 방식 | 문자의 OCR 변환 작업으로 구축되는 DB 형태 |
| | 한적 자료 | 고문서, 고도서 등과 같이 한자로만 쓰여진 자료를 입력 작업으로 구축되는 DB 형태 |
| 이미지 | | 스캐닝 또는 카메라 촬영을 통하여 구축되는 DB 형태 |
| 사운드 | | 녹음 또는 보유자료(tape)의 편집으로 구축되는 DB 형태 |
| 동영상 | | 촬영 또는 보유자료(reel tape, 베타 tape, 비디오 tape)의 편집으로 구축되는 DB 형태 |
| 3D | | 디지털 촬영을 통하여 나온 이미지를 3차원 데이터로 구축하는 이미지 기반 모델링 및 렌더링 방식과 3D 스캐닝을 통해 3차원 데이터로 구축되는 DB 형태 |
| GIS | | 기 제작된 지도의 스캐닝 및 속성 정보를 입력 등으로 구축되는 DB 형태 |
| 항공사진 | | 필름 및 사진형태로 보관되어 있는 항공사진에 촬영정보 및 공간정보를 수록하여 구축되는 DB 형태 |
| 기상위성사진 | | 과거 위성원시자료 및 지구관측위성 이진자료를 표준 포맷으로 전환하여 구축되는 DB 형태 |
| 지도제작위성사진 | | 위성사진에 속성정보를 입력하고 수치정사영상 자료로 구축되는 DB 형태 |

〈표 2-5〉, 〈표 2-6〉, 〈표 2-7〉, 〈표 2-8〉은 각각 동영상과 이미지, 사운드, GIS 유형에 적용 가능한 품질기준을 정의한 사례이다.

1) 한국정보화진흥원(NIA), "데이터베이스 구축 방법론 Ver.3", 2006.

〈표 2-5〉 동영상에 대한 품질기준 정의 사례

| 주특성 | 정 의 | 부특성 | 상세 내용 |
|---|---|---|---|
| 기능성 (Functionality) | 해당 콘텐츠가 특정 조건에서 사용될 때, 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 정도 | 적절성 (Suitablility) | 목적에 대한 내용의 부합여부, 운용상의 적절성(사이즈, 한 프레임의 크기, 파일 포맷, 초당 프레임수, 초당전송율, 압축율, 화면비율, Runinng Time 등) 등 |
| | | 정확성(Accuracy) | 규격에 따른 구현의 정확성 (메타데이터 정확성 및 메타데이터와의 일치 여부) |
| | | 상호운용성 (Interoperability) | 동기화(Synchronization) - 사운드 동기화, 자막동기화 등 |
| | | 기능순응성 (Functionality Compliance) | 구축/품질 표준에 대한 준수여부 (명명규칙, 기준값 준수 등) |
| 신뢰성 (Reliability) | 해당 콘텐츠가 규정된 조건에서 사용될 때 규정된 신뢰수준을 유지하거나 사용자로 하여금 오류를 방지할 수 있도록 하는 정도 | 성숙성(Maturity ) | 기준 환경(모니터, 스피커, 컴퓨터 최소 사양)에서 결함발생 정도 |
| | | 신뢰순응성 (Reliability Compliance) | 신뢰성 관련한 표준의 준수여부 |
| 사용성 (Usability) | 해당 콘텐츠가 규정된 조건에서 사용될 때, 사용자에 의해 이해되고, 선호될 수 있게 하는 정도 | 이해성 (Understandability) | 영상 끊김, 영상인식 만족정도, 음향인식 만족정도 등 |
| | | 친밀성 (Attractiveness) | 사용자에게 친숙하고 사용이 용이한 포맷 사용여부 |
| | | 사용순응성 (Usability Compliance) | 사용성 관련 표준 준수여부 |
| 효율성 (Efficiency) | 해당 콘텐츠가 규정된 조건에서 사용되는 자원의 양에 따라 요구된 성능을 제공하는 정도 | 시간효율성 (Time Behaviour) | Loading Time의 적절성 |
| | | 자원효율성 (Resource Behaviour) | 기준 환경(모니터, 스피커, 컴퓨터 최소 사양)의 적절성 |
| | | 효율순응성 (Efficiency Compliance) | 효율성과 관련된 표준의 준수여부 |
| 이식성 (Portability) | 해당 콘텐츠가 다양한 환경과 상황에서 실행될 수 있는 가능성의 정도 | 적응성 (Adaptability) | 운영환경 및 플레이어 호환성 |
| | | 공존성 (Co-Existence) | Running 시 다른 S/W의 동작에 영향 여부 |
| | | 이식순응성 (Portability Compliance) | 이식성과 관련된 표준을 준수하는 정도 |

〈표 2-6〉 이미지에 대한 품질기준 정의 사례

| 주특성 | 정의 | 부특성 | 상세 내용 |
|---|---|---|---|
| 기능성 (Functionality) | 해당 콘텐츠가 특정 조건에서 사용될 때, 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 정도 | 적절성 (Suitablility) | 목적에 대한 내용의 부합여부, 운용상의 적절성(사이즈, 파일 포맷, 해상도, 압축율 등), 저장방법(비트맵(bitmap) 방식 또는 래스터(raster) 방식, 벡터(vector) 방식) 등 |
| | | 정확성 (Accuracy) | 규격에 따른 구현의 정확성 (메타데이터 정확성 및 메타데이터와의 일치 여부) |
| | | 기능순응성 (Functionality Compliance) | 구축/품질 표준에 대한 준수여부 (명명규칙, 기준값 준수 등) |
| 신뢰성 (Reliability) | 해당 콘텐츠가 규정된 조건에서 사용될 때 규정된 신뢰수준을 유지하거나 사용자로 하여금 오류를 방지할 수 있도록 하는 정도 | 성숙성 (Maturity ) | 기준 환경(모니터, 스피커, 컴퓨터 최소사양)에서 결함발생 정도 |
| | | 신뢰순응성 (Reliability Compliance) | 신뢰성 관련한 표준의 준수여부 |
| 사용성 (Usability) | 해당 콘텐츠가 규정된 조건에서 사용될 때, 사용자에 의해 이해되고, 선호될 수 있게 하는 정도 | 이해성 (Understand ability) | 이미지 인식 만족정도 |
| | | 친밀성 (Attractiveness) | 사용자에게 친숙하고 사용이 용이한 포맷 사용여부 |
| | | 사용순응성 (Usability Compliance) | 사용성 관련 표준 준수여부 |
| 효율성 (Efficiency) | 해당 콘텐츠가 규정된 조건에서 사용되는 자원의 양에 따라 요구된 성능을 제공하는 정도 | 시간효율성 (Time Behaviour) | Loading Time의 적절성 |
| | | 자원효율성(Resource Behaviour) | 기준 환경(모니터, 스피커, 컴퓨터 최소사양)의 적절성 |
| | | 효율순응성 (Efficiency Compliance) | 효율성과 관련된 표준의 준수여부 |
| 이식성 (Portability) | 해당 콘텐츠가 다양한 환경과 상황에서 실행될 수 있는 가능성의 정도 | 적응성 (Adaptability) | 운영환경 및 Viewer호환성 |
| | | 공존성 (Co-Existence) | Viewing 시 다른 S/W의 동작에 영향 여부 |
| | | 이식순응성 (Portability Compliance) | 이식성과 관련된 표준을 준수하는 정도 |

〈표 2-7〉 사운드에 대한 품질기준 정의 사례

| 주특성 | 정의 | 부특성 | 상세 내용 |
|---|---|---|---|
| 기능성 (Functionality) | 해당 콘텐츠가 특정 조건에서 사용될 때, 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 정도 | 적절성 (Suitablility) | 목적에 대한 내용의 부합여부, 운용상의 적절성(사이즈, 파일 포맷, 음질, 압축율, 기본 소리크기, Runinng Time 등) 등 |
| | | 정확성 (Accuracy) | 규격에 따른 구현의 정확성 (메타데이터 정확성 및 메타데이터와의 일치 여부) |
| | | 기능순응성 (Functionality Compliance) | 구축/품질 표준에 대한 준수여부 (명명규칙, 기준값 준수 등) |
| 신뢰성 (Reliability) | 해당 콘텐츠가 규정된 조건에서 사용될 때 규정된 신뢰수준을 유지하거나 사용자로 하여금 오류를 방지할 수 있도록 하는 정도 | 성숙성(Maturity) | 기준 환경(모니터, 스피커, 컴퓨터 최소사양)에서 결함발생 정도 |
| | | 신뢰순응성 (Reliability Compliance) | 신뢰성 관련한 표준의 준수여부 |
| 사용성 (Usability) | 해당 콘텐츠가 규정된 조건에서 사용될 때, 사용자에 의해 이해되고, 선호될 수 있게 하는 정도 | 이해성 (Understand ability) | 음향인식 만족정도(듣고 이해할 수 있는 정도) 등 |
| | | 친밀성 (Attractiveness) | 사용자에게 친숙하고 사용이 용이한 포맷 사용여부 |
| | | 사용순응성 (Usability Compliance) | 사용성 관련 표준 준수여부 |
| 효율성 (Efficiency) | 해당 콘텐츠가 규정된 조건에서 사용되는 자원의 양에 따라 요구된 성능을 제공하는 정도 | 시간효율성 (Time Behaviour) | Loading Time의 적절성 |
| | | 자원효율성 (Resource Behaviour) | 기준 환경(모니터, 스피커, 컴퓨터 최소사양)의 적절성 |
| | | 효율순응성 (Efficiency Compliance) | 효율성과 관련된 표준의 준수여부 |
| 이식성 (Portability) | 해당 콘텐츠가 다양한 환경과 상황에서 실행될 수 있는 가능성의 정도 | 적응성 (Adaptability) | 운영환경 및 플레이어 호환성 |
| | | 공존성 (Co-Existence) | Running 시 다른 S/W의 동작에 영향 여부 |
| | | 이식순응성 (Portability Compliance) | 이식성과 관련된 표준을 준수하는 정도 |

〈표 2-8〉 GIS에 대한 품질기준 정의 사례

| 주특성 | 정의 | 부특성 | 상세 내용 |
|---|---|---|---|
| 기능성 (Functionality) | 해당 콘텐츠가 특정 조건에서 사용될 때, 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 정도 | 완전성 (Completeness) | • DB 내에 누락의 에러가 없음을 의미함<br>• 데이터완전성(data completeness) : DB가 표준 및 규약에 기술된 객체들을 모두 포함함<br>• 모델완전성(model completeness) : DB에 적용된 표준 및 규약이 특정 응용시스템에 적합함 |
| | | 해상도 (Resolution) | • 시간, 공간, 주제 상에서 구분할 수 있는 데이터의 상세한 정도를 의미함<br>• 어떠한 측정시스템도 완벽하게 정확하지 않고, DB 용량 문제로 상세함을 다소 희생해야 하는 한계가 있음.<br>• 해상도 수준은 정확도에 영향을 준다. |
| | | 정확성 (Accuracy) | • 에러에 반대되는 개념으로, EAV entity-attribute-value)모델을 기반으로 정의<br>• entity는 실세계 공간사상을, attribute는 관련된 특성(property)을, value는 속성의 질적 또는 양적 측면임<br>• 공간(spatial), 시점(temporal), 주제(thematic) 측면에서의 정확도로 측정됨 |
| | | 일관성 (Consistency) | • 한 DB 내에서의 모순의 유무를 의미함<br>• 다양한 데이터 출처, 인력, 장비 등에 따른 공간적 편차가 없는 것 |
| | | 준수성 (Compliance) | 관련 표준 및 규약에 대한 준수 정도 |

## 2. 품질 이슈 조사

### 2.1 개요

데이터 품질 이슈 조사는 업무 담당자와 면담을 통하여 비즈니스적 관점에서 담당자가 인식하고 있는 저품질의 문제점을 발견하기 위한 기법이다. 사전에 수립된 면담 계획에 따라 업무 담당자와 면담을 실시하여 정보시스템의 정형·비정형 데이터에 대한 신뢰 수준과 데이터의 정확성과 관련된 이용자의 만족도를 조사하고 부정확한 데이터를 발생시키는 주요 현상을 파악한다.

### 2.2 수행절차

1) 데이터 품질 이슈조사에 활용할 수 있는 수집 가능한 자료 및 문서를 체계적으로 수집한 후 취합하여 내용을 파악한다. 수집할 주요 자료 및 문서는 다음과 같다.

- 과거 수행 품질진단 프로젝트 관련자료
- 데이터 아키텍처 자료
- 과거 데이터 변경 및 클린징 자료
- 장애 유형 분석 자료

2) 수집한 자료를 토대로 면담 질의서를 작성한다. 현 이슈조사의 목적이 주요 업무 데이터 항목이나 주요 비정형 데이터를 파악하고 해당 데이터의 주요 오류발생 현상을 분석하는 것이므로, 주요 데이터 항목 또는 비정형 데이터 규명·데이터의 오류발생 유형·시점·시스템 등의 현상 파악에 기초하여 면담 질의서를 작성한다. 면담 질의서 작성 시 다음 사항을 고려한다.

- 주요 시스템 및 업무데이터 항목 또는 비정형 데이터 파악
- 정보시스템의 데이터 신뢰 수준
- 부정확한 데이터를 발생시키는 주요 요인
  - 데이터 입력 및 변경 시 또는 비정형 데이터 작성 시 오류 발생
  - 데이터 누락 현황
  - 데이터 구축 및 통합 이후 지속적으로 발생되는 오류 데이터 현상
  - 데이터 추출·변환·탑재·통합 등의 데이터 이동 및 재구조화시의 불일치 발생 유형
- 데이터의 표준화 수준 및 요구사항
- 데이터의 정확성과 관련된 이용자 만족

3) 사전에 수립된 면담 계획에 따라 해당 업무 담당자와 면담을 실시한다. 면담 주제는 수행 작업과 대상자의 담당 업무를 고려하여 결정한다. 또한 면담을 실시하기 전에 면담 일정, 장소, 면담요지를 면담 대상자에게 사전 배포하여 답변 내용이나 관련 자료를 미리 준비할 수 있도록 한다.

4) 조사된 면담결과를 분석하여 이슈보고서를 작성하여 면담 대상자에게 확인을 받는다.

### 2.3 산출물 : 데이터 품질 이슈 보고서

## 3. 데이터 관리 문서 수집

### 3.1 개요

정형·비정형 데이터에 대한 품질진단을 수행하기 전에 데이터 규칙이나 표준 사항을 사전에 파악해야 한다. 데이터 규칙은 데이터 관리 문서의 분석이나 업무담당자의 업무 지식으로부터 파악될 수 있고, 데이터 관련 표준은 업무 담당자나 조직의 품질 담당자로부터 확인할 수 있어야 한다. 특히 비정형 데이터에 대해서는 작성 표준이 사전에 확정되고 이에 따라 데이터 또는 콘텐츠가 작성되어야 하기 때문에 비정형 데이터에 대한 품질진단에 있어서 표준 문서의 확인은 필수적이다.
데이터 규칙을 파악하기 위해 수집해야 할 주요 문서는 데이터 구축 프로젝트에서 산출물로 정의되고 있는 데이터 관리 산출물이다. 예를 들어 전체 테이블 명세서・주요 업무테이블 현황・컬럼 정의서・도메인 정의서・표준코드 테이블・엔티티 관계도・업무 메뉴얼 등이 해당된다.

### 3.2 수행절차

데이터 규칙과 표준을 파악하기 위해 필요한 데이터 관리 문서를 사전에 정의한다. 데이터 관리 문서에는 데이터 관리 규정, 모델, 데이터 표준, 콘텐츠 작성 표준, 데이터베이스 운영과 관련된 문서 등이 있다.

〈표 2-9〉 데이터 규칙 파악에 필요한 문서 예시

| 구 분 | 문서 내용 | 필수 |
|---|---|---|
| 데이터모델 | ERD(개념, 논리, 물리 모델) | √ |
| | 엔티티 정의 및 속성 정의 | √ |
| | 테이블 정의 및 컬럼 명세 | √ |
| 데이터표준 | 명명표준(엔티티, 테이블, 속성, 컬럼, Tablespace명, Index명, constraint명) | √ |
| | 콘텐츠 작성 표준 (동영상, 이미지, 사운드 등 비정형 콘텐츠에 해당) | √ |
| | 용어사전 | √ |
| | 코드목록 및 코드정의 | √ |
| | 도메인 정의 | √ |
| 데이터베이스 | 시스템 목록 | √ |
| | 시스템 구성도 | √ |
| | 시스템 운영 방안 및 관련 자료 | √ |
| | … | … |

2) 사전 정의된 데이터 관리문서 목록을 IT 부서에 요청하여 수집한다. 수집대상 문서 형식은 가급적 워드・HWP・엑셀・ERWIN 파일 등의 전자화된 문서를 수집한다. 수집한 문서를 품질진단 수행 범위・대상시스템・관련 업무별로 분류한다.

3) 수집된 관리 문서들 중에서 테이블 및 컬럼 정의서, ERD, 관계정의서 등을 토대로 문서를 분류하여 테이블 및 컬럼 목록과 관계 목록을 작성한다. 문서 포맷은 엑셀이나 스프레드쉬트 등을 활용한다. 테이블 컬럼 목록과 관계 목록은 향후 품질진단 시 정확한 데이터 규칙을 규명하기 위한 용도로 활용된다.

3.3 산출물 : 데이터 관리문서, 테이블 및 컬럼, 관계 목록

### 3.4 참고사항

앞서 언급한 관리 문서 이외에도 자료흐름도, 상태변화도, 응용프로그램 소스나 프로그램 명세서 등이 필요할 경우 IT부서에 요청하여 수집한다. 또한 업무 담당자가 알고 있는 업무 지식은 업무 담당자와의 면담을 통하여 파악한다.

## 4. 진단 대상 중요도 평가

### 4.1 개요

조직에서 운영하는 정형·비정형 데이터는 관리 목적에 따라 평가되는 중요도가 서로 상이하다. 예를 들어 고객 데이터는 데이터 운영 조직에서 매우 중요하게 여기고 있지만 데이터 항목별 활용 목적에 따라 중요도에 차이가 생길 수 있다. 예컨대 우편물을 취급하는 우체국이나 택배회사의 경우에는 고객의 주소가 매우 중요하나 텔레마케팅 회사에서는 조사에 활용할 고객의 전화번호가 더욱 중요하다. 따라서 데이터를 평가하기 위한 중요도를 업무 및 데이터 항목별로 사전에 선정하여 평가해야 된다. 이는 비정형 데이터에 대해서도 동일하게 적용되는데, 예를 들어 GIS 데이터의 경우 어떤 업무나 사용 목적인지에 따라 요구되는 정밀도는 다를 수 있으며, 이에 따라 데이터 정밀도에 대한 중요도는 달라질 수 있다.
평가된 중요도는 향후 품질진단을 수행할 대상 테이블과 컬럼을 선정하거나 향후 핵심 정보항목을 도출하기 위한 용도, 또는 진단 대상 콘텐츠와 이에 따른 측정 항목을 선정하기 위한 용도로 활용되며, 품질측정 단계에서 항목별 오류율 배점에 가중치의 차등을 두기 위한 목적으로 활용된다.

## 4.2 수행절차

1) 테이블 및 컬럼 또는 비정형 콘텐츠 및 해당 측정 항목의 중요도를 선정하기 위하여 사전에 중요도 평가기준을 설계한다. 중요도 평가기준은 대상 시스템과 수록된 데이터를 관리하고 활용하는 조직의 관점에 따라 다양하게 도출될 수 있으므로 별도의 품질위원회를 소집하여 각 분야 전문가와 업무 담당자들의 의견을 최대한 반영하는 것이 바람직하다.

2) 설계된 평가기준과 중요도를 기준으로 진단 대상 테이블 및 컬럼, 또는 비정형 콘텐츠 및 해당 측정 항목을 평가한다. 사전에 취합한 테이블 및 컬럼 목록, 또는 비정형 콘텐츠 및 해당 측정 항목을 업무 전문가들에게 배포하여 평가한다. 다수의 업무 전문가를 평가위원으로 선정하여 중요도를 평가할 경우, 각 평가된 중요도의 적절한 대표값(최빈치 선정 또는 평균)을 선정함으로써 평가된 중요도의 객관성을 높일 수 있다.

3) 평가된 중요도 목록을 취합한 후 사전 정의된 중요도 계산 원칙에 따라 종합 중요도를 계산하여 기입한다. 취합된 테이블 및 컬럼, 또는 비정형 콘텐츠 및 해당 측정 항목의 중요도 목록을 업무전문가들에게 리뷰를 거친 후 최종 확정한다.

4) 진단 대상에 따라 서로 다른 품질기준 또는 품질기준의 가중치를 다르게 적용할 수 있다. 진단 대상에 대한 품질기준 및 품질기준의 가중치를 평가하는 경우 계층화분석

법(AHP, Analytic Hierarchy Process)과 같은 검증된 가중치 도출 방법을 사용하거나 사전에 설계된 평가기준에 의거하여 평가할 수 있다.

4.3 산출물 : 진단 대상 중요도 평가기준 정의서, 진단 대상 중요도 목록

## 5. 품질진단 대상 선정

### 5.1 개요

데이터 품질진단을 수행할 대상 테이블 및 컬럼, 또는 비정형 콘텐츠 및 해당 측정 항목을 선정한다. 진단 대상 목록은 향후 품질측정 및 모니터링의 대상이 됨과 동시에 지속적으로 관리해야 할 중요 데이터 항목으로 선정되므로 업무 전문가와 데이터 품질 담당자가 공동으로 대상 선정 작업을 수행한다.

### 5.2 수행절차

1) 사전에 수집한 품질 이슈 조사서, 중요도 평가 결과를 수집하여 품질진단 대상 선정 목록을 작성하고, 객관적인 선정기준을 작성한다.

2) 진단 대상 선정기준은 사전에 수행된 데이터 항목별 중요도 산정기준, 품질 이슈 대상 항목 선정 기준과 일관되게 작성한다.

3) 진단 대상 선정 기준 및 목록을 업무 담당자에게 배포하

여 진단 대상을 선정하도록 요청한다.

4) 배포된 목록을 취합하여 품질진단 대상으로 확정하고, 별도의 검토회의를 거쳐서 확정한다.

5.3 산출물 : 품질진단 대상 목록, 진단 대상 선정 기준

## 6. 핵심 품질 항목 선정

### 6.1 개요

핵심 품질 항목은 회사의 고객, 업무 프로세스, 재무 환경 등에서 직접으로 영향을 미치는 중요성이 매우 높은 데이터를 말한다. 이러한 데이터들은 조직 내의 데이터 품질기준과 매핑되어 관리된다.

### 6.2 수행절차

1) 사전에 수집된 진단 대상 항목을 기준으로 주요 데이터 품질 이슈 정보항목과 주요 업무 데이터를 기준으로 도출 대상 후보를 작성하고 해당 핵심 데이터 후보 항목별 중요도를 정의한다.

2) 도출된 핵심 데이터 후보 항목을 품질전문가와 업무전문가들이 심의 위원회를 구성하여 업무 영향도, 우선순위, 적용 가능여부, 적정성 등을 평가하여 최종 핵심 품질 항목으로 결정한다.

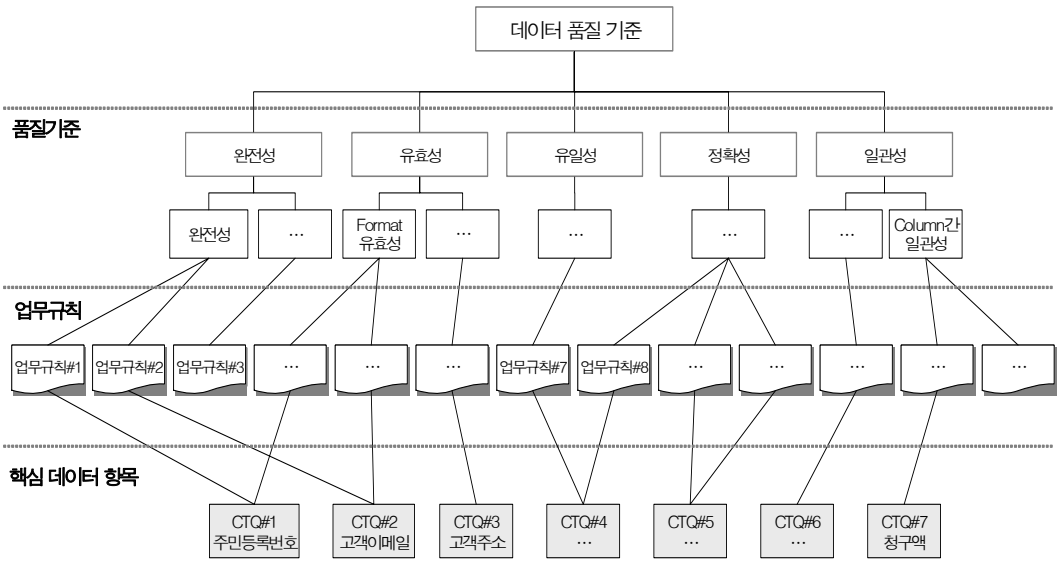
〈표 2-10〉 핵심 품질 항목 도출 예시

| 업무 | 영문 테이블명 | 한글 테이블명 | 영문 컬럼명 | 한글 컬럼명 | 중요도 | | | | | | 이슈 대상 | 진단 대상 | 핵심 데이터 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 마스터 데이터 | 컬럼 중요도 | 시스템 연관성 | 서비스 영향도 | 업무 영향도 | 중요도 | | | |
| 사업 관리 | PROJECT | 사업 | PRJ_NO | 사업번호 | A | A | A | A | A | 3.0 | | √ | √ |
| | PROJECT | 사업 | PRJ_NM | 사업명 | A | A | A | A | A | 3.0 | √ | √ | √ |
| | PROJECT | 사업 | DEPT_CD | 사업부처 | A | A | A | A | A | 3.0 | √ | √ | √ |
| | PROJECT | 사업 | START_YY | 사업시작년도 | B | A | B | B | B | 2.7 | √ | √ | |
| | PROJECT | 사업 | END_YY | 사업종료년도 | B | A | B | B | B | 2.7 | | √ | |
| 간행물 관리 | PAPER | 논문 | KIND_GB | 학술지구분 | A | A | A | A | A | 3.0 | | | |
| | PAPER | 논문 | ANNNC_DT | 논문발표일 | A | B | A | A | A | 2.3 | √ | | √ |
| | PAPER | 논문 | TITLE | 제목 | A | A | A | A | A | 3.0 | | √ | √ |
| | PAPER | 논문 | VOL_NM | 게게권/집 | C | B | C | C | C | 1.7 | √ | | |
| | PAPER | 논문 | NUM_NO | 게재호 | C | B | C | C | C | 1.7 | | √ | |
| | PAPER | 논문 | START_PG_NO | 시작페이지 | C | C | C | C | C | 1.0 | | | |
| | PAPER | 논문 | END_PG_NO | 종료페이지 | C | C | C | C | C | 1.0 | | | |

## 6.3 산출물 : 핵심 품질 항목 정의서

## 6.4 참고사항

핵심 품질 항목은 지속적인 관리가 필요하며, 별도의 관리 주기를 정하여 업무 변화에 따른 지속적인 변경 이력을 관리한다. 핵심 품질 항목은 향후 품질측정 시 업무규칙 도출 기준과 매핑하여 핵심 데이터 항목 별로 품질 수준을 측정하기 위한 용도로 활용되며 조직에서 관리해야 할 우선 품질 개선 대상으로 활용한다.

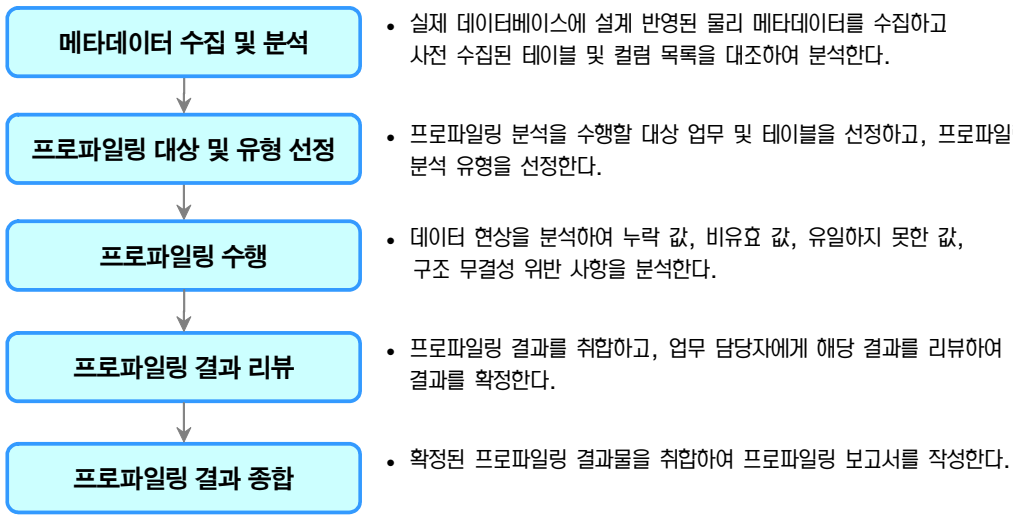



〈그림 2-4〉 데이터 품질기준 및 핵심 데이터 항목 활용

## 7. 데이터 프로파일링

데이터 프로파일링은 주로 정형 텍스트 데이터 및 비정형 콘텐츠의 메타데이터에 대한 품질진단에 활용되며, 통계적 기법을 활용하여 데이터의 품질과 관련된 현상을 파악하는 절차로서 데이터 소스에 존재하는 데이터의 구조, 내용, 품질을 파악하기 위해 다양한 형태로 분석하는 절차이다. 다시 말해 데이터에 대한 정보를 추출하는 것이다.

데이터 프로파일링은 메타데이터와 대상 소스데이터에 대한 통계적 분석 결과를 통해 데이터 품질 문제를 이슈화하고 개선점을 찾는 것을 주된 목적으로 한다.




〈그림 2-5〉 데이터 프로파일링 수행 절차

## 7.1 메타데이터 수집 및 분석

### 7.1.1 개요

수집해야 할 주요 메타데이터는 실제 운영 중인 데이터베이스의 테이블명·컬럼명·자료형·도메인·제약조건 등이며 데이터베이스 설계에는 반영되지 않은 한글 메타데이터·도메인정보·엔티티 관계·코드정의 등도 포함한다. 올바른 메타데이터의 수집은 향후 데이터 프로파일링 분석을 수행할 때 컬럼 속성과 구조 정보를 규명하는데 기초 자료가 되므로 매우 중요하다.

### 7.1.2 수행절차

1) 사전에 수집한 데이터 관리문서·테이블 목록·컬럼 목록·

관계 목록을 준비한다.

2) 데이터 소스에 접속하여 물리 데이터베이스에 존재하는 컬럼 레이아웃 또는 카탈로그 정보를 추출한다.

3) 추출된 컬럼 레이아웃 정보와 데이터 관리 문서를 매핑하여 불일치 사항을 분석한다. 관리문서에 누락되거나 불일치하는 테이블 및 컬럼명, 각 자료형 및 길이 등을 분석하고 표준화 원칙에 위배되는 사항 등을 분석하여 검토의견을 작성한다.

4) 분석된 메타 정보 및 검토의견을 취합하여 정리한다.

7.1.3 산출물 : 메타데이터 불일치 테이블 목록 및 컬럼 목록

## 7.2 프로파일링 대상 및 유형 선정

### 7.2.1 개요

데이터 프로파일링을 수행할 대상 테이블 및 컬럼 그리고 프로파일링 유형을 사전에 선정한다. 데이터 프로파일링은 실제 운영데이터의 소스에 접속하여 대상 데이터 항목에 대한 분석을 실시하므로 프로파일링 분석 유형과 데이터 항목의 수록된 레코드 수량에 따라 분석 소요시간, 측정방법이 서로 상이하다. 따라서 분석 가능한 데이터 항목과 유형을 사전에 선정한 후 프로파일링을 수행하는 것이 바람직하다.

### 7.2.2 수행절차

1) 사전에 작성한 테이블 및 컬럼 목록, 관계 목록을 토대로 해당 컬럼별, 측정 유형별 프로파일링 대상 목록을 작성한다. 프로파일링 수행 대상 선정 시 다음 사항을 고려한다.
   - 컬럼의 레코드 건수
   - 컬럼의 필수 여부 및 기본값 설계 여부(NOT NULL, DEFAULT)
   - 컬럼의 유형(키컬럼, 식별자성 컬럼)
   - 컬럼의 포맷(날짜유형, 코드화 형식)
   - 컬럼의 도메인
   - 프로파일링의 샘플링 비율

2) 수집된 데이터 관리 문서를 토대로 프로파일링을 수행할 대상 유형과 해당 컬럼을 확인한다. 구조분석 대상의 경우에는 통합코드와 관계목록을 구분하여 제약사항을 기재한다.

### 7.2.3 산출물 : 데이터 프로파일링 수행 대상 목록

### 7.2.4 참고사항

데이터 품질과 관련된 이슈사항이 없거나 초기 데이터 품질 현상 파악을 위해 데이터 프로파일링을 수행하는 경우에는 데이터의 결함을 가장 잘 파악할 수 있는 컬럼을 대상으로 프로파일링을 수행하는 것이 효율적이다.
다음은 프로파일링을 수행할 대상 테이블명 및 컬럼명, 프로파일링 유형을 나열한 예제이다.

〈표 2-11〉 프로파일링 대상 및 유형 목록

| 영문 테이블명 | 한글 테이블명 | 영문 컬럼명 | 한글 컬럼명 | 데이터 타입 | 길이 | Null | 초기값 | 프로파일링 유형 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 완전성 | 유효성 | | | 유일성 | … |
| | | | | | | | | | 날짜 | 범위 | 포맷 | | |
| PROJECT | 사업 | PRJ_NO | 사업번호 | VARCHAR2 | 8 | N | | √ | | | √ | √ | |
| PROJECT | 사업 | PRJ_NM | 사업명 | VARCHAR2 | 255 | N | | √ | | | | | |
| PROJECT | 사업 | DEPT_CD | 사업부처 | VARCHAR2 | 4 | N | | | | | √ | | |
| PROJECT | 사업 | START_YY | 사업시작 년도 | CHAR | 4 | N | | | √ | | | | |
| PROJECT | 사업 | END_YY | 사업종료 년도 | CHAR | 4 | N | | | √ | | | | |
| PAPER | 논문 | KIND_GB | 학술 지구분 | VARCHAR2 | 4 | N | | √ | | | √ | | |
| PAPER | 논문 | ANNNC_ DT | 논문 발표일 | CHAR (10) | | Y | | | √ | | | | |
| PAPER | 논문 | VOL_NM | 게재권/집 | VARCHAR2 | 128 | Y | | | | | | | |
| PAPER | 논문 | NUM_NO | 게재호 | NUMBER | 11 | Y | | | | √ | | | |
| PAPER | 논문 | START_ PG_NO | 시작 페이지 | NUMBER | 11 | N | 0 | | | √ | | | |
| PAPER | 논문 | END_PG_ NO | 종료 페이지 | NUMBER | 11 | N | 0 | | | √ | | | |

다음은 구조 분석을 위한 대상 테이블 및 컬럼 관계 목록을 정리한 예제이다. 관계를 측정해야 할 기준 테이블 및 컬럼명과 대상 테이블 및 컬럼명을 나열하고, 관계명, 기수성 등을 나열한다.

〈표 2-12〉 구조 분석 대상 목록

| 구분 | 기준 테이블명 | 기준 컬럼명 | 대상 테이블명 | 대상 컬럼명 | 대상 키컬럼명 | 관계명 | null 허용 | 분석 대상 | 비 고 |
|---|---|---|---|---|---|---|---|---|---|
| 관계 | MEMBER | MEMBER_NO | MEMBER_DTL | MEMBER_NO | MEMBER_NO | RI_01 | N | √ | 1:M |
| | PROJECT | PRJ_NO | PROJECT_SPLY | PRJ_NO | SPLY_ID | - | N | √ | 1:M |
| | PRODUCT | PRODUCT_CD | REPAIR_DTL | PRODUCT_CD | REPAIR_SN | - | N | √ | |
| | CUSTOMER | CUSTOMER_ID | ORDER | CUSTOMER_ID | ORDER_ID | RI_03 | N | √ | |
| 코드 | CODE_MASTER | CD | PROJECT | DEPT_CD | PROJECT_ID | | Y | √ | TYPE_CD = '003' |
| | CODE_MASTER | CD | PAPER | KIND_GB | PAPAER_ID | | N | √ | TYPE_CD = '004' |
| | CODE_MASTER | CD | SUPPLIER | WRK_GB | SUPPLIER_NO | | Y | √ | TYPE_CD = '014' |
| | CODE_MASTER | CD | SUPPORT | SPRT_GB | SUPPORT_NO | | N | √ | TYPE_CD = '027' |
| | CODE_MASTER | CD | ORDER | CMPN_GB | ORDER_NO | | N | √ | TYPE_CD = '033' |

## 7.3 데이터 프로파일링 수행

### 7.3.1 개요

사전 선정된 대상 테이블 및 컬럼, 분석 유형별로 프로파일링을 수행한다. 본 지침서에서는 완전성 분석·유일성 분석·유효성 분석·일관성 분석의 4가지 품질기준 영역에서 수행할 수 있는 프로파일링 분석 유형을 제안한다. 자세한 프로파일링 분석 기법은 3장을 참조한다.

### 7.3.2 수행절차

1) 사전 정의된 프로파일링 대상 테이블 및 컬럼에 대하여 기초 통계 값 분석을 실시하여 값의 현상을 추출한다. 다음은 데이터의 오류 규칙을 발견하기 위하여 사용되는 주요 집계 특성을 나타낸다.

- 널 값의 분포
- 공백 값 및 0 값의 분포
- 값의 범위(최대값, 최소값)
- 값의 출현 빈도 범위(최대 빈도 값, 최소 빈도 값)
- 입력된 값의 개별 값 목록과 건수
- 문자의 패턴
- 유일한 값의 갯수
- 값의 평균, 분산, 표준편차

2) 수집된 통계치를 분석하여 해당 품질기준별 오류 발생 현상을 파악하고, 결과를 정리한다. 수집된 통계치 분석을 통하여 직관적으로 파악되는 누락 값·공백 값·무의미한 값 등을 토대로 완전성·유일성·유효성 등의 위반 내역을 파악한다.

3) 분석된 컬럼 중에서 오류 데이터가 발견되어 정밀한 검사가 요구되는 컬럼과 별도의 유효성 분석을 실시할 대상을 중심으로 추가 유효성 분석을 실시한다. 추가 유효성 분석은 유효해야 할 값의 목록·범위·날짜 유형·패턴 등을 미리 정의하여 검사하는 것으로서 일반적인 프로파일링 도구에서 기본 기능으로 제공한다. 프로파일링 도구가

구비되지 않은 경우에는 별도의 SQL문 또는 유효성 검증 함수를 작성하여 값의 유효성 위반 내역을 확인한다.

4) 유효값 목록 이외의 관계목록, 통합코드 목록은 관계 중첩도(포함도) 분석 함수를 적용하여 구조 무결성 위반 내역을 분석한다.

5) 분석한 유효성 위반 컬럼과 해당 컬럼의 규칙 위반 레코드 건수・해당 컬럼의 유효값 목록・유효범위・유효패턴・유효 날짜유형・관계 위반 목록 등을 취합한다.

### 7.3.3 산출물 : 프로파일링 분석서

## 7.4 프로파일링 결과 취합 및 리뷰

### 7.4.1 개요

프로파일링 분석이 완료되면 분석 유형별 오류발생 컬럼에 대하여 담당자의 리뷰과정을 수행한다. 담당자가 오류로 확인한 프로파일링 결과물을 취합하여 프로파일링 분석 결과서를 작성한다.

### 7.4.2 수행절차

1) 프로파일링 분석 결과를 측정 유형별로 정리하여 컬럼 속성의 오류발생 유형을 취합하여 프로파일링 분석서를 목록으로 작성한다.

2) 작성된 프로파일링 분석서를 초기 품질진단 계획 수립 시 배정된 업무 담당자에게 전달하여 오류데이터발생 컬럼의 진위 확인을 요청한다. 해당 업무 담당자가 오류발생 컬럼에 대하여 대한 이견이 있을 경우 담당자의 검토 의견을 작성하도록 요청한다.

3) 검토 완료된 프로파일링 분석서들을 각 해당 업무 담당자들로부터 회수하여 오류데이터 발생 컬럼을 해당 업무별, 테이블별, 오류발생 유형별로 정리하여 최종 프로파일링 결과 종합 보고서를 작성하여 해당 업무 담당자에게 배포한다.

7.4.4 산출물 : 프로파일링 분석 결과서

## 8. 업무규칙 도출

업무규칙(Business Rule)은 주로 정형 텍스트 데이터 및 비정형 콘텐츠의 메타데이터에 대해 적용되는 것으로, 조직이 목표를 수행하는데 적용할 수 있는 운영·정의·제약사항의 기술 규칙이다. 업무규칙은 비즈니스의 구조를 설명하거나, 제어 또는 업무 및 어떠한 조치 사항에 영향을 가하기 위한 용도로 활용된다.

업무규칙은 비즈니스 관점의 업무규칙과 정보시스템 관점의 업무규칙으로 구분된다. 비즈니스 관점의 업무규칙은 전사적 또는 개념적 모델로 접근하며 정보시스템 관점의 업무규칙은 논리적 모델로 접근한다. 예를 들어 “우리 회사는 신용 등급이 나쁜 고객에게는 어떠한 제품도 판매하지 않는다는 원칙을 준수한다.”는 비즈니스관점의 업무규칙이다. “만일 고객의 신용등급이 'C'미만일 경우에는 구매주문 레코드는 입력되어서는 안된다.”는 정보시스템 관점의 업무규칙의 사례이다.

정보시스템 관점의 업무규칙은 속성규칙·구조규칙·유도값 규칙·프로세스 규칙으로 다시 구분할 수 있다. 정보시스템의 데이터 값 생성·변경·삭제 원칙이 업무 로직과 일치해야 하는데 정확해야 하는 모든 규칙을 업무규칙으로 정의한다. 업무규칙은 시스템을 운영하는 조직의 업무관리 규정 및 정책에 따라 결정되며, 업무규칙 구성요소로는 조직의 정책, 자연언어로 기술된 업무규칙 및 업무규칙 설명, 정형화된 규칙 문장, 정형화 표현 형식 등이 있다.

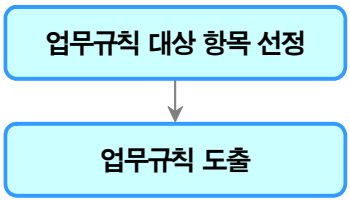
속성 규칙은 해당 속성의 값이 준수해야 할 규칙으로서 단일 속성의 기본값 · NULL · 유일성 규칙 · 해당 도메인 · 유효범위 · 유효값 목록 · 자료 포맷 · 자료형 · 제약사항 등을 의미한다.

구조 규칙은 모델의 구조상 준수해야 할 규칙으로서 함수적 종속관계에 있는 속성간의 일관성, 다수 테이블간 관계에 있는 속성 간의 참조무결성 및 관계별 기수성(Cardinality) 등을 의미한다. 공통코드 테이블 및 마스터 테이블을 참조하는 이력, 거래내역, 트랜잭션 테이블은 참조하는 마스터 테이블의 식별자에 등록된 값을 사용해야 된다는 규칙이 이에 해당된다.

유도값 규칙은 타 데이터 항목 또는 다수의 데이터 항목의 조합·계산·집계 등으로 특정 컬럼의 속성값이 유도되어 결정될 경우 해당 속성 값이 준수해야 할 규칙을 의미한다. 일반적으로 나이 · 재직기간 · 매출총액 · 통계테이블 등과 같이 다수의 컬럼의 계산 및 조합으로 유도된 컬럼의 정확성 확보를 위한 규칙이 이에 해당된다.

프로세스 규칙은 업무 프로세스의 변화에 따라 데이터 값이 준수해야 되거나 값의 제약사항으로 인해 업무 프로세스에 제약사항을 주는 규칙을 의미한다. 예를 들어 “신용등급이 ‘CCC’미만인 고객은 대출을 제한한다”라는 규정이 있을 경우, 값 관점에서는 “고객의 신용등급이 CCC일 경우 대출테이블의

대출일, 대출금액은 NULL이어야 한다.”라는 규칙이 존재한다. 또한 프로세스 관점에서는 “고객의 신용등급이 'CCC'일 경우에는 대출일, 대출금액 등이 시스템에서 입력되지 못하도록 제한을 가하여, 업무 담당자는 신용 등급이 낮은 고객에게 대출업무를 수행하지 못하게 하는 제약사항을 가할 수 있다. 업무규칙의 작성은 일관되고 정형화된 양식으로 작성한다. 또한 품질기준별 오류발생 추이를 분석하기 위해 업무규칙에 해당 데이터 품질기준과 핵심데이터 여부를 명시한다.




- 프로파일링결과, 품질 이슈조사 결과서를 토대로 업무규칙 도출 대상 항목을 선정한다.
- 선정된 업무규칙을 도출 대상 항목을 기준으로 업무규칙 정의서를 작성한다.

〈그림 2-6〉 업무규칙 도출 절차

## 8.1 업무규칙 도출 대상 항목 선정

### 8.1.1 개요

업무규칙은 지속적으로 유지 관리해야 할 데이터와 관련된 규칙이므로 사전에 수행된 프로파일링 결과・품질이슈・핵심데이터 항목 등을 토대로 조직 내부에서 품질 개선을 위해 중점적으로 관리해야 할 대상을 선별하여 이를 바탕으로 도출하며 지속적으로 관리해야 된다. 따라서 사전에 수행한 프로파일링 수행 결과와 진단 대상 목록, 핵심데이터 등의 자료를 토대로 업무규칙을 도출할 대상 데이터 항목을 사전에 정의하고 이를 토대로 업무규칙을 도출한다.

### 8.1.2 수행절차

1) 데이터 관리문서, 테이블 및 컬럼 중요도, 프로파일링 결과서, 품질이슈 정의서, 핵심 데이터 항목 정의서 등 사전에 수행된 작업을 토대로 자료를 수집한다.
2) 수집된 진단 대상 테이블 및 컬럼 목록과 프로파일링 결과서, 품질 이슈 결과서를 취합하여 업무규칙 도출 대상 목록을 작성한다.
3) 작성된 대상 목록을 토대로 업무규칙 도출 대상을 선정한다. 프로파일링 결과 오류가 발생하는 대상을 위주로 해당 테이블 및 컬럼의 중요도, 핵심정보항목 등을 고려하여 업무규칙 도출 순서를 정하여 업무규칙 도출 대상을 선정한다.
4) 업무규칙을 적용할 데이터 항목별 측정 기준·용도·주기 핵심 데이터 여부·해당 중요도 등을 결정한다.
5) 도출된 대상 데이터 항목 중에서 누락되었거나 중복된 데이터 항목이 있는지 점검한다. 단일 데이터 항목이 자료사전과 공통코드 테이블에 정의되었는지 재차 점검하고 누락되거나 중복된 사항이 발견되면 정정한다.

### 8.1.3 산출물 : 업무규칙 후보 목록

### 8.1.4 참고사항

프로파일링 단계에서 파악된 오류발생 컬럼은 단일 컬럼 규칙 위반 내역 또는 구조적 무 결성에 위배되는 오류발생 데이터 항목이다. 일반적으로 업무규칙에는 값, 구조 이외의 업무 프로세스와 연관되어 발생될 수 있는 복잡한 규칙이 존재한다. 이러한 복잡한 업무규칙은 프로파일링 단계에서 발견

되기 어려우며 업무규칙 작성 단계에서 업무 담당자가 또는 품질 분석가가 별도의 규칙으로 작성한다.

## 8.2 업무규칙 도출

### 8.2.1 개요

사전에 준비된 업무규칙 도출 대상 항목으로부터 해당 품질기준과 중요도, 핵심 정보항목과 매핑하여 업무규칙 정의서를 작성한다. 업무규칙은 데이터의 품질관리를 위해 지속적으로 관리되어야 하는 데이터의 규칙으로 데이터의 값이 정확하기 위한 조건에 대한 표현이다. 따라서 업무규칙은 일관되고 정형화된 양식으로 작성된다.

### 8.2.2 수행절차

1) 사전에 작성된 업무규칙 도출 대상 항목을 기준으로 업무규칙 정의서를 작성한다. 업무규칙은 해당 품질기준과 사전에 파악되었던 데이터 항목별 중요도, 핵심 데이터 항목 여부 등을 기재한다.

2) 작성된 업무규칙의 적정성을 평가한다. 업무규칙은 품질분석가와 업무 담당자들이 공동 작업으로 도출하므로 도출된 업무규칙의 표현이 일관되지 못하거나 작업자간의 견해 차이로 적절하지 못한 업무규칙이 도출 될 수 있다. 업무규칙 보완 작업 시 확인할 사항은 다음과 같다.
   - 중요 업무규칙의 누락이 없는지 파악한다.
   - 업무규칙의 중복성을 제거한다.
   - 업무규칙이 정확히 도출되었는지 파악한다.

- 업무규칙 표현이 업무규칙 작성 표준과 일치하고 일관된 표현을 사용하는지 파악한다.
- 가독성이 떨어지거나 난해한 표현으로 기술되었는지 파악한다.

3) 보완된 업무규칙 정의서를 모두 취합하여 확정한다. 확정된 업무규칙은 결재권자의 승인을 얻어 공식 문서화 한다. 업무규칙은 데이터의 개정, 변경 등의 사유로 해당 규칙의 변경이 요구될 때 그 개정 및 변경 이력을 모두 관리한다.

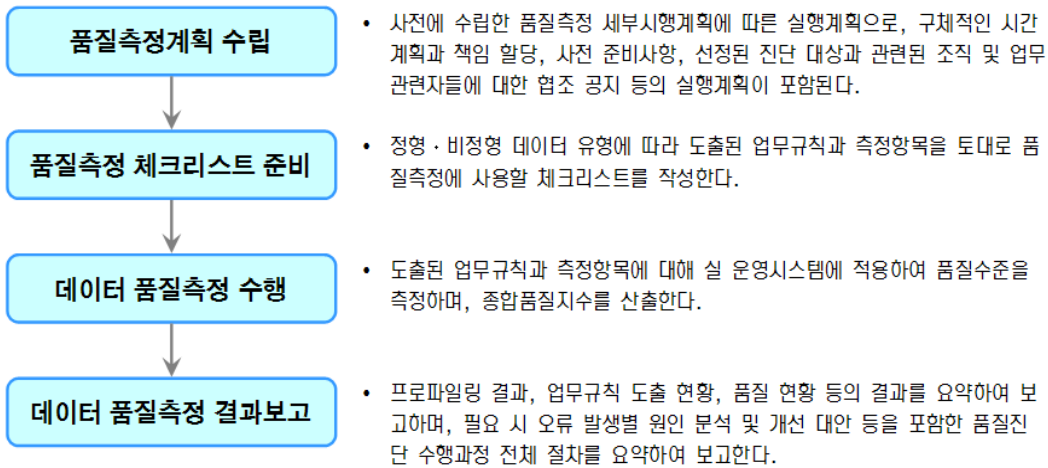
### 8.2.3 산출물 : 업무규칙 정의서

### 8.2.4 참고사항

업무 담당자가 SQL로 구성된 측정 스크립트를 직접 작성하기 어려울 수 있으므로 업무 담당자는 해당 업무규칙명과 업무규칙에 대한 상세 설명을 작성하고 작성된 사항을 토대로 IT 부서 담당자나 품질전문가가 측정 스크립트를 작성하여 기입할 수 있다.

## 제3절 데이터 품질측정

데이터 품질측정 단계는 도출된 업무규칙을 핵심데이터에 적용하여, 오류데이터를 추출하고 그 현황을 파악하는 오류데이터 검증 단계와 그 결과를 요약하여 품질 현황을 보고하는 단계로 구성된다.




〈그림 2-7〉 데이터 품질측정 절차

## 1. 품질측정 계획 수립

### 1.1 개요

품질측정 계획은 사전에 수립한 품질측정 세부 시행 계획에 따른 상세 실행계획이다. 품질측정을 위한 상세 절차에 대해 구체적인 시간 계획을 할당하고, 각 절차에서의 수행자와 책임사항을 정의한다. 품질측정 전반 및 각 절차에서의 사전 준비사항 등에 대해 파악하고, 선정된 진단 대상과 관련된 조직 및 업무 관련자들에게 공지할 협조사항 등을 정의한다.

### 1.2 수행절차

1) 데이터 품질을 측정하기 위한 상세 절차 또는 수행 내용을 구분한다.

2) 구분된 각 상세 절차에 대해 수행 시간, 장소, 수행자, 사전 준비사항, 책임사항 등을 정의한다.

3) 선정된 진단 대상과 관련된 조직 및 업무 관련자들을 확인하고 공지할 내용이 있는 경우 이에 대한 공지계획(업무협조공문, 사내방송, 게시판 공지, 관련 사이트 팝업창 공지 등 공지방법과 공지내용, 공지기간, 공지책임자 등을 포함)을 수립한다.

4) 수립된 상세실행계획은 품질측정에 참여할 관련자들과 리뷰하여 내용을 보완하고 공유한다.

### 1.3 산출물 : 품질측정 상세실행계획

### 1.4 참고사항

업무규칙 도출 이후 초기 데이터 품질측정 시에는 업무규칙의 측정 주기와 장소, 시기 등을 준수하지 않고 특정 시점에서 일괄로 적용할 수 있다. 업무규칙의 측정은 프로파일링과 다르게 샘플링을 실시하지 않으므로 운영 중인 데이터 레코드 전수를 대상으로 품질을 측정할 경우 운영시스템에 과부하를 줄 수 있다. 이러한 경우는 업무 외의 시간이나 시스템의 유휴 시간을 활용하여 데이터 품질을 측정하는 것이 바람직하다. 이러한 상황을 파악하고 고려하여 구체적인 실행계획을 수립하는 것이 품질측정 계획 수립 단계이다.

## 2. 품질측정 체크리스트 준비

### 2.1 개요

정형·비정형 데이터 유형에 따라 도출된 업무규칙과 측정항목을 토대로 품질측정에 사용할 체크리스트를 작성한다. 이

를 통해 진단 대상에 대해 측정할 업무규칙이나 측정항목이 누락됨이 없이 측정에 사용될 수 있도록 한다. 품질측정은 체크리스트에 수록된 내용을 토대로 수행된다. 품질측정 결과 역시 체크리스트에 기록되어 오류율 및 품질지수 산출에 사용된다.

### 2.2 수행절차

1) 정형 텍스트 데이터의 경우 데이터 품질을 측정할 대상 테이블 및 컬럼별로 해당 업무규칙의 목록을 준비한다. 비정형 콘텐츠의 경우는 콘텐츠 유형별로 선정된 측정항목의 목록을 준비하고, 각 측정항목에 대해 측정내용을 작성한다.

2) 준비된 업무규칙이나 측정항목의 목록에 대해 중요도 평가 내용, 측정 총건수, 오류 건수, 오류율, 가중오류율 등 측정 결과를 기록하기 위한 항목을 추가하여 체크리스트를 완성한다.

**2.3 산출물** : 품질측정 체크리스트

**2.4 참고사항** : 체크리스트에 대한 상세 내용 및 작성 사례는 3, 4장에서 소개한다.

## 3. 데이터 품질 측정 수행

### 3.1 개요

정형·비정형 데이터 유형별로 사전에 도출된 업무규칙이나 측정항목별 측정내용을 실제 운영 데이터베이스에 적용하여

오류데이터 및 오류율을 측정하여 업무, 품질기준 및 핵심정보항목별 오류율 및 품질 지수를 산출한다.

### 3.2 수행절차

1) 정형·비정형 데이터 유형별로 사전에 도출된 업무규칙이나 측정항목별 측정내용이 수록된 품질측정 체크리스트를 토대로 수록 내용을 실제 운영 중인 물리 데이터베이스 및 콘텐츠에 적용하여, 해당 데이터 항목이 지켜야할 규칙에 위반되는 오류데이터 발생 내역과 오류건수를 추출한다.

2) 추출된 데이터의 오류현황과 오류내역을 업무규칙별·품질기준별·핵심정보항목별로 취합하여 목록을 작성하며 해당 업무규칙별 중요도나 품질기준 항목에 따라 품질 지수를 산출한다.

**3.3 산출물** : 업무규칙별 오류율 현황, 핵심데이터별 오류율 현황, 데이터 유형별 품질지수 현황

### 3.4 참고사항

품질측정은 사전에 수립된 품질측정계획에 따라 정해진 시간, 장소, 대상에 대해 수행하여 품질측정이 해당 조직의 업무수행에 대해 지장을 초래하지 않도록 해야 한다.

## 4. 데이터 품질측정 결과 보고

### 4.1 개요

프로파일링 결과, 업무규칙 도출 현황, 콘텐츠 유형별 측정 항목 및 측정 내용 도출 현황, 품질측정 결과 등의 내용을 요약하여 관련자들에게 알려주게 되며, 필요 시 오류 발생별 원인 분석 및 개선 방안 등이 포함된 품질진단 전체 수행과정 및 절차를 요약하여 종합적인 보고를 수행할 수도 있다. 일반적으로 오류 원인별 개선 방안은 다음 단계에서 품질개선계획 수립 시 사용된다.

### 4.2 수행절차

1) 측정 결과에 대해 해당 업무규칙별 중요도나 품질기준 항목에 따라 품질 지수를 산출하여 업무규칙별 · 품질기준별 · 핵심정보항목별 품질 현황자료를 취합한다.

2) 품질측정 결과 보고서를 작성하여 해당 업무전문가와 업무 담당자에게 전달한다.

### 4.3 산출물 : 품질측정 결과 보고서

### 4.4 참고사항

필요에 따라 오류 원인 분석 및 개선 방안 수립을 선행하여 이를 포함한 종합 보고서를 작성할 수 있다.

## 5. 데이터 품질 종합 보고서 작성

### 5.1 개요

상세하고 종합적인 보고가 필요하다고 판단되는 경우 진단 데이터 대상·프로파일링 분석·업무규칙·품질측정·오류 발생 유형별 원인 분석 등 품질진단 수행 과정의 요약 및 현황, 주요 이슈 사항을 요약하여 종합 보고서를 작성한다. 작성된 품질 종합 보고서는 해당 조직의 품질관리 의사결정자에게 보고된 후 향후 데이터 품질 개선 업무에 활용된다.

### 5.5 수행절차

1) 진단 보고서의 목차, 내용 및 범위를 결정한다. 데이터 품질진단 결과를 활용할 목적과 이용할 대상자를 명확히 하고, 진단 보고서에 포함시킬 내용과 범위 등을 결정한다. 진단 목적, 진단 대상, 진단 방법 및 절차, 진단 결과 등을 포함한다.

2) 도출된 목차, 내용, 프로젝트의 범위를 감안하여 평가 보고서를 작성한다.
   - 진단 계획
     - 진단 목적 및 범위
     - 진단 유형
     - 진단 대상
     - 진단 일정
   - 진단 경과
     - 진단 방법
     - 진단 조직 및 투입 인력
   - 진단 결과

- 진단 유형별 품질측정 결과
- 품질 수준 및 품질 지수
- 주요 오류발생 유형 및 원인 분석

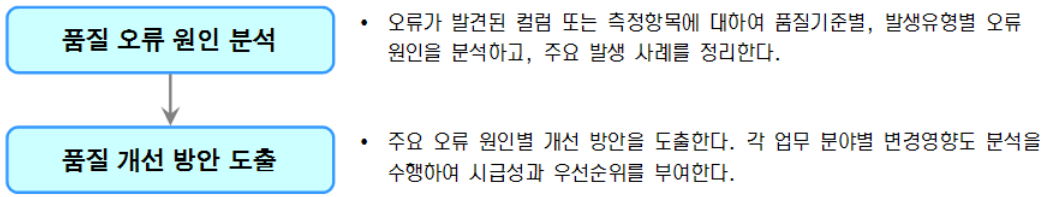
• 진단 산출물

3) 작성된 진단 보고서를 검토한다. 진단 보고서의 검토를 위하여 초안을 준비하고, 업무 전문가에게 사전에 배포하여 검토를 의뢰한 후에 검토회의를 거쳐서 최종 확정한다.

5.6 **산출물** : 데이터 품질 종합 보고서

## 제4절 데이터 품질측정 결과 분석

데이터 품질측정이 완료되면 오류 유형에 따른 발생 원인을 분석하고 이에 따라 각 업무 분야에 해당되는 개선방안을 도출한다.




〈그림 2-8〉 데이터 품질측정 결과 분석 절차

## 1. 품질오류 원인 분석

### 1.1 개요

데이터의 품질측정이 완료되면 주요 오류데이터와 오류 발생 유형을 분석하여 주요 오류발생 원인을 규명해야 된다. 따라서 오류원인 분석 절차에서는 데이터 표준화·데이터구조·데이터 입력 및 흐름 통제·응용프로그램의 오류 등의 전체 시스템 운영 관점에서 오류 발생 현상을 해석하고 분석한다. 오류 원인분석은 각 발생 유형별로 상세하게 기술하여 품질 개선활동 전에 명확히 전개한다. 정형 텍스트 데이터의 경우 오류데이터 내역과 업무규칙을 근거로 해당 오류의 발생 주기와 원인을 확인하고 분석한다.

### 1.2 수행절차

1) 품질측정 단계에서 얻어진 데이터 품질 현황 및 업무별·핵심 데이터항목별·품질기준별 오류발생 자료를 준비하고 해당 오류발생 현황을 파악한다.

2) 오류가 발생하는 업무규칙(정형 데이터 및 비정형 콘텐츠의 메타데이터 경우)이나 측정항목(비정형 콘텐츠의 경우)의 중요도, 오류율 등을 고려하여 우선순위를 선정하고 데이터의 오류발생 원인분석 작업을 준비한다. 주요 원인분석 대상으로 선정된 업무규칙이나 측정항목의 중요도 및 시급성을 고려하여 해당 업무 담당자와 원인분석 일정 계획을 협의한다.

3) 원인 분석 대상이 되는 업무규칙이나 측정항목에 대하여 데이터 품질측정결과 및 업무규칙 정의서 등을 토대로 오

류데이터 발생 주기 및 경향을 파악한다.

- 오류발생 주기 : 지속적 발생 혹은 특정시점에서 발생
- 오류발생 원인 : 응용프로그램 결함・입수데이터 결함・과거 오류 데이터・데이터 보정 작업 결함・원인 불명 등

4) 오류를 발생 시키는 주요 원인을 추적하여 각 유형을 분석한다. 오류발생 원인이 응용프로그램일 경우 오류가 발생된 테이블 및 컬럼과 해당 응용프로그램간의 CRUD Matrix[2] 상관성 분석을 실시하여 해당 응용프로그램의 오류 원인과 프로그램 모듈을 추적한다. 또한 오류가 발생하는 테이블과 컬럼에 대한 DB-응용프로그램간 Alignment 분석[3], 물리 데이터모델-DB Alignment[4] 분석을 수행하여 서로 불일치하는 내역을 정리한다. 오류발생 원인이 데이터 흐름으로부터 발생된 경우 소스 대 타깃 매핑 분석과 소스 데이터 추출 모니터링 보고서 등의 검토를 통하여 그 원인이 흐름과정에서 추출・변환・적재 조건의 오류에 기인한 것인지 검토한다. 비정형 콘텐츠의 경우 작성 표준의 부재에 따른 품질 불균일 문제인지, 작성 도구의 문제인지, 작성자의 기술수준에 따른 문제인지, 또는 불명확한 요구정의에 따른 문제인지 등의 여부도 검토한다.

5) 오류가 발생된 업무규칙이나 측정항목의 유형에 따라 원인을 확정하여 그 내용을 상세히 정리한다. 해당 업무규칙이나 측정항목을 데이터 품질 개선 대상 리스트로 취합

2) CRUD 매트릭스 분석은 응용프로그램의 기본 프로세스의 액션(Create:생성 Read:조회 Update:수정 Delete:삭제)과 해당 관련 업무 테이블과의 매트릭스를 작성하여 서로의 상관성을 분석하는 기법으로, 어떤 프로그램이 어던 테이블을 접근하는가를 분석할 수 있으며, 데이터 모델과 프로세스 모델의 적합성을 판단할 수 있다.

3) DB와 응용프로그램 사이의 차이를 분석하는 방법을 일컫는다.

4) 물리 데이터 모델이 DB에 적절히 반영되었는가의 차이를 분석하는 방법을 일컫는다.

하여 오류원인 분석서 및 조치 계획서를 준비한다.

1.3 산출물 : 오류 원인 분석서

## 2. 품질 개선 방안 도출

### 2.1 개요

오류 원인분석 수행 결과에 따라 그 해당되는 정형·비정형 데이터, 저장 구조, 응용프로그램의 모든 영역에서의 개선점을 찾는다. 따라서 품질 개선안을 도출하기 이전에 각 분야별 변경 영향도 분석을 사전에 수행한다. 도출된 개선사항은 시급성과 우선순위에 따라 순차적으로 수행한다.

### 2.2 수행절차

1) 표준데이터, 모델데이터, 정형·비정형 데이터와 응용프로그램에 대한 각각의 변경 영향도 검토 과정을 거쳐 오류가 발생된 데이터의 변경에 따른 예상되는 개선방안과 그 세부 수행 계획을 도출한다.

2) 개선방안에는 데이터 정제 또는 변경·재작성(비정형 콘텐츠의 경우), 표준 변경, 구조 변경, 응응 프로그램 변경, 데이터 흐름 통제 계획 변경 등이 포함되며, 각 개선사항에 대한 위험 분석·타당성 검토·각 중요도·예상 변경 영향도 등을 종합적으로 평가하여 그 우선순위를 부여한다.

3) 사전 도출된 오류원인 분석사항과 개선사항을 토대로 조치 계획서를 작성한다. 조치계획서에 오류발생 주기·오류원인 분석·개선 및 조치사항 등을 상세히 기록한다.

데이터의 정제가 필요할 경우에는 정제와 관련된 사항을 반드시 기록한다.

4) 업무 담당자들과 함께 사전 도출된 오류원인 분석결과와 개선방안에 대한을 검토회의를 소집하여 개선방안을 확정하여 공식화한다. 공식화된 개선사항은 관련 향후 업무 담당자들에게 배포되며 각 분야의 업무전문가는 배포된 개선방안을 토대로 개선활동을 수행한다.

**2.3 산출물** : 변경영향도 분석서, 오류원인분석 및 조치계획서

## 2.4 참고사항

〈표 2-13〉 변경영향도 분석서 예시

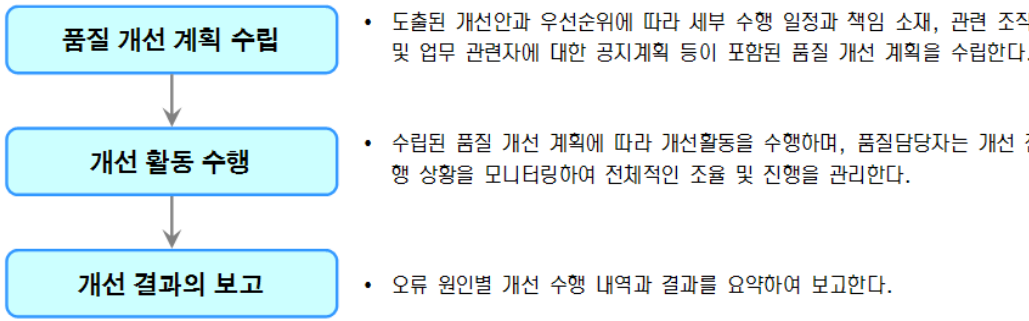
<table>
<tr><th colspan="10">데이터 변경영향도 평가서</th></tr>
<tr><td rowspan="4">변경항목</td><td>시스템명</td><td>DB명</td><td>테이블명</td><td>컬럼명</td><td>변경구분</td><td colspan="4">변경사항</td></tr>
<tr><td>고객관리</td><td>RQMS</td><td>CUST_<br>MST</td><td>CUST_CD</td><td>코드변경</td><td colspan="4">CUST_CD의 자료형을 CHAR (6)에서 VARCHAR(10)로 변경하되 기존데이터 앞에 '0000'을 채울 것</td></tr>
<tr><td>담당부서</td><td>고객<br>관리부</td><td>직위</td><td>대리</td><td>내선</td><td>101</td><td>성명</td><td colspan="2">김표준</td></tr>
<tr><td>변경사유</td><td colspan="8">1) 초기 시스템 설계시 고객 코드를 너무 작게 설계<br>2) 현재 고객수가 80만명 초과</td></tr>
<tr><td>우선순위</td><td colspan="3">매우 시급</td><td>중요도</td><td colspan="4">매우 중요</td></tr>
<tr><td rowspan="6">관련항목</td><td>시스템명</td><td>DB명</td><td>테이블명</td><td>컬럼명</td><td>담당부서</td><td>직위</td><td>내선</td><td>담당자</td><td>변경영향</td></tr>
<tr><td>주문관리</td><td>RQMS</td><td>GD_ORDR</td><td>CUST_CD</td><td>주문관리부</td><td>대리</td><td>403</td><td>김주문</td><td>200만개 레코드</td></tr>
<tr><td>결재관리</td><td>RQMS</td><td>GD_SNCT</td><td>CUST_CD</td><td>주문관리부</td><td>대리</td><td>404</td><td>이결재</td><td>120만개 레코드</td></tr>
<tr><td>배송관리</td><td>RSMS</td><td>GD_SND</td><td>CUST_CD</td><td>물류서비스부</td><td>과장</td><td>501</td><td>박배송</td><td>90만개 레코드</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td colspan="9"></td></tr>
<tr><td>조치사항</td><td colspan="9">• 중요 테이블(마스터)의 설계 변경<br>• 48개 컬럼과 연관되어 있으며, 우선순위는 매우 시급함<br>• 레코드 증가추이 분석결과 2008년 4월 이후 100만 회원을 초과할 예정이므로 늦어도 2008년 2월까지는 변경이 완료되어야 함</td></tr>
</table>

〈표 2-14〉 오류원인분석 및 조치계획서 예시

| 오류원인 분석 및 조치계획서 | | | | | |
|---|---|---|---|---|---|
| 업무규칙ID | BR20080105 | | | | |
| 업무구분 | 과제관리 | 업무명 | 과제심사 | | |
| 업무규칙 | 심사절차코드가 심사중(011)일 경우에는 심사일은 반드시 있어야 한다. | | | | |
| 데이터베이스명 | 과제DB | | | | |
| 관련테이블명 | PRJ_DTL | 관련컬럼명 | PRJ_DTL.STTS<br>PRJ_DTL.STTS.INSP_DT, | | |
| 품질기준 | 완전성 | 세부품질기준 | 조건완전성 | | |
| 오류발생 주기 | □ 항상 발생<br>■ 특정시점에서 발생<br>– 발생시기 : 저장을 하였을 경우, 트랜잭션 오류로 심사일의 입력이 누락되는 경우가 발생 | | | | |
| 오류데이터 유형 | • 입력통제오류<br>• 조건 완전성 오류 | | | | |
| 오류 원인 분석 | • 응용프로그램 오류<br>• 한 화면에서 다수의 트랜잭션을 처리 | | | | |
| 개선 및 조치사항 | • 과제관리〉 과제심사관리 〉 심사처리 모듈 수정 요망<br>• 데이터 정제 필요 | | | | |
| 정제 관련사항 | • 심사절차코드가 심사중(011)이며, 심사일이 누락된 경우, 해당 레코드의 심사일을 심사 종료일로 업데이트할 것 | | | | |
| 작성자 | 김정제 | 작성일자 | 2008.01.01 | 조치 예정일 | 2008.01.31 |

# 제5절 데이터 품질 개선

데이터 품질진단과 오류 원인 분석, 개선방안 도출이 완료되면 오류유형과 각 업무 분야에 해당되는 개선방안의 시급성과 우선순위에 따라 데이터 품질 개선계획을 수립하고 개선 활동을 수행한다. 개선 활동이 완료되면 개선 결과를 검토하고 요약하여 보고한다.




〈그림 2-9〉 데이터 품질 개선 절차

## 1. 품질 개선 계획 수립

### 1.1 개요

오류 원인분석 수행 결과에 따라 품질 개선안이 도출되면 시급성과 우선순위에 따라 순차적으로 수행하기 위한 실행 계획을 수립한다.

### 1.2 수행절차

1) 업무 분야별 개선방안의 시급성과 우선순위에 따라 순차적 수행이 가능하도록 개선안들을 나열한다.

2) 개선활동을 수행할 기간과 파급효과, 관련 조직 또는 업무 관련자 등을 고려하여 각 개선안별로 실행 일정과 책임자 등을 할당한다.

3) 개선안의 내용에 따라 관련 조직이나 업무관련자의 업무 수행에 영향이 미치는 경우는 해당 내용을 공지하기 위한 방법과 공지내용, 공지기간, 공지책임자 등을 정의한다.

4) 품질개선계획 수립이 완료되면 개선활동에 참여할 IT부서 및 관련 업무 전문가들과 함께 검토회의를 하여 계획 내용을 확정하여 공식화한다.

1.3 산출물 : 품질개선수행계획서

## 2. 개선 활동의 수행

### 2.1 개요

사전에 수립된 품질개선수행계획에 따라 개선 방안의 중요도, 우선순위 및 업무 영향도 등을 감안하여 개선활동을 수행한다. 수행된 활동은 개선 평가 과정을 통하여 해당 실적 평가를 실시한 후 지속적인 품질 확보를 위한 관리 계획을 수립한다.

### 2.2 수행절차

1) 사전에 도출된 개선 방안별 중요도와 우선순위에 따른 세부 수행 일정을 확인하고, 승인된 개선 방안을 토대로 개선활동을 실시한다.

2) 품질담당자는 품질개선수행계획에 따라 각 개선안들이 일정대로 원활하게 수행되고 있는지, 예상치 못한 이슈사항이 발생했는지 등을 지속적으로 모니터링하여 개선활동을 관리한다.

3) 개선활동이 완료되면 초기 품질개선 계획 수립 시 정량화하여 제시되었던 측정 목표치 대비 개선된 사항의 목표치를 측정해야 하므로 사전에 진행했던 품질의 측정 및 분석 절차를 재차 수행하여 데이터 품질과 관련된 개선 실적을 평가한다.

2.3 산출물 : 품질 개선 결과서

## 3. 개선 결과의 보고

### 3.1 개요

개선활동이 완료되면 개선결과를 평가하고 지속적인 수행이 필요한지 여부에 따라 지속계획을 수립하는 한편, 미개선이 있는 경우는 그 사유를 파악하여 개선활동 보고서를 작성한다.

### 3.2 수행절차

1) 개선활동이 완료된 이후 개선 사항과 미개선 사항을 정리하여 데이터 품질 개선 활동 보고서를 작성한다. 개선활동 보고서에는 개선 일정·활동 내역·미개선 사유 등의 내용을 포함한다.

2) 사전에 진행했던 품질측정 및 분석 절차를 재차 수행하여

평가된 개선 실적은 데이터 품질 개선활동 보고서 등을 통하여 품질관리 책임자나 경영진에 보고한다.

3) 수행되었던 데이터 품질 현황·품질 개선활동·모니터링 계획·품질진단 프로젝트 수행 시 발생된 품질 이슈사항 등을 정리하여 향후 데이터 품질확보 및 관리계획을 수립한다.

### 3.6 산출물 : 품질 개선활동 보고서

### 3.7 참고사항

경영진은 지속적인 데이터 품질저하를 예방하기 위하여 데이터 품질관리를 위한 자원을 재배분하거나, 품질관리체계를 수립하는 등 조직 관점의 데이터 품질 향상을 위한 관리 정책을 수립해야 한다. 지속적 품질관리를 위한 고려사항은 다음과 같다.

- 데이터 품질관리 체계의 도입
  - 품질관리 조직의 구성 및 편성
  - 데이터 품질관리 프로세스 개선
- 전사 데이터 표준화 및 구조 관리
  - 전사 표준화 정책 도입으로 일관성 확보
  - 업무 담당자간 의사소통의 원활한 진행
  - 통합코드를 일관되게 관리하여 코드의 중복사용과 불명확성 제거
  - 효율적인 업무 협의 지원
  - 코드 오류의 발생을 사전 예방

- 데이터 모델 현행화 및 확산으로 안정성 확보

• 품질관리 솔루션의 도입 검토
  - 표준 및 메타데이터관리 솔루션·품질진단 도구·데이터의 흐름 통제도구·데이터 클린징 솔루션 등의 도입 검토
  - 업무 프로세스 자동화로 데이터 관리 업무의 편의성 제공
  - 데이터 값·구조·관리 업무의 최신성 유지
  - 정보흐름을 파악하고 오류발생 가능성을 미연 예방
• 데이터 품질관리 인증 검토
  - 현 데이터 품질관리 활동의 수준을 측정
  - 자체평가, 감독자에 의한 평가, 컨설팅, 전문기구에 의한 인증 검토

# 제3장 데이터 품질진단 기법 - 정형 데이터

본 장에서는 정형 텍스트 데이터 및 비정형 콘텐츠의 메타데이터 등에 적용할 수 있는 품질진단 기법에 대해 기술한다. 데이터 프로파일링 기법과 업무규칙 도출 사례를 중심으로 정형 텍스트 데이터의 데이터 품질진단 기법에 관하여 기술한다. 또한 데이터 품질측정 스크립트를 기준으로 분석 유형별 프로파일링 기법 및 업무규칙 도출 방법을 소개하고 해당 오류데이터 검증 SQL의 사례를 소개한다.

## 제1절 데이터 프로파일링

데이터프로파일링(Data profiling)은 데이터 소스에 대해 일련의 데이터 검사 절차를 수행함으로써 데이터에 관한 중요한 정보와 통계치를 수집하는 것이다. 이는 데이터베이스에 있는 방대한 정보로부터 숨어있는 지식(hidden knowledge)을 자동적으로 추출하는 과정인 데이터 마이닝 혹은 지식 발견의 개념과 유사하다. 나아가 데이터 프로파일링은 데이터의 구조·내

용·품질을 발견하기 위해 다양한 분석 기술을 활용한다.

현재 운영되는 시스템에 수록된 데이터의 오류발생 현상을 파악하기 위하여 데이터 프로파일링 기법은 매우 합리적인 해결책이다. 데이터 프로파일링은 데이터의 통계적 분석 방법을 적용하여 시스템에 적재된 데이터와 관련된 오류 현상을 발견을 할 수 있고 이렇게 발견된 현상을 토대로 관리 문서와 시스템간의 불일치 사항을 제안하여 고품질 데이터를 을 할 수 있는 기반 환경을 제공할 수 있다.

데이터 프로파일링은 발견(Discovery)과 검증(Verification)이라는 절차로 구성된다. 데이터의 발견의 절차를 통하여 오류의 가능성이 있는 부정확한 데이터 현상이 발견되고 발견된 현상은 관련 업무 담당자들과 품질 분석가의 협의를 거쳐서 부정확한 데이터로 결정된다. 결정된 부정확한 데이터 규칙과 정확한 데이터 규칙을 토대로 오류데이터 내역을 추출하거나 검증하는 행위를 하는데 이 행위를 데이터의 검증이라 한다. 일부 조직에서는 데이터를 검증하는 행위를 데이터 감사(Data Auditing)라고도 한다.

프로파일링 기법을 통해 추출된 오류데이터와 데이터 규칙은 향후 데이터 현행화라는 품질개선 활동의 기초 작업이 되는 동시에 향후 품질관리 활동의 핵심 업무 영역이 된다.

## 1. 메타데이터 수집 및 분석

### 1.1 메타데이터 수집

메타데이터는 데이터를 위한 데이터이다. 메타데이터의 수집은 프로파일링을 수행하기 이전 단계에서 수행되어야 하며 데이터의 부정확성을 판단하는데 매우 중요한 기초 자료가 되므로 가능한 모든 정확한 메타데이터를 수집해야 된다.

메타데이터의 수집은 데이터 관리 문서로부터 수집한다. 데

이터 품질진단을 수행하기 전에 데이터 관리문서를 사전 정의하여 모두 수집하고 확보해야 된다. 수집해야할 주요 데이터 관리 문서는 테이블 정의서・컬럼 정의서・도메인 정의서・데이터 사전・ERD・관계 정의서 등이 있다.

데이터 품질진단 시 기초 자료로 활용할 테이블 및 컬럼 정의서・ERD・관계정의서 등을 토대로 테이블 및 컬럼 목록과 관계 목록을 표 형식으로 작성한다.

〈표 3-1〉 테이블 및 컬럼 목록 예시

| 영문 테이블명 | 한글 테이블명 | 영문 컬럼명 | 한글 컬럼명 | 데이터 타입 | 길이 | Null | 초기값 | 비고 |
|---|---|---|---|---|---|---|---|---|
| PROJECT | 사업기본정보 | PRJ_NO | 사업번호 | VARCHAR2 | 8 | N | | |
| PROJECT | 사업기본정보 | PRJ_NM | 사업명 | VARCHAR2 | 255 | N | | |
| PROJECT | 사업기본정보 | PRJ_DPT | 사업부처 | VARCHAR2 | 4 | N | | |
| PROJECT | 사업기본정보 | PRJ_STTDT | 사업시작년도 | CHAR | 4 | N | | |
| PROJECT | 사업기본정보 | PRJ_ENDDT | 사업종료년도 | CHAR | 4 | N | | |
| FUNDING | 투자정보 | FND_NO | 투자번호 | VARCHAR2 | 8 | N | | |

테이블 관계 목록은 관계를 가지는 테이블과 컬럼 목록, 관계의 기수성, 그 외의 제약사항과 통합코드 테이블일 경우는 코드 구분 등을 명시하여 작성한다.

〈표 3-2〉 관계 및 코드 목록 예시

| NO | 구분 | 부모 테이블명 | 주요키 컬럼명(PK) | 분석 대상 테이블명 | 분석 대상 컬럼명(FK) | 관계명 | 기수성 | 조건 |
|---|---|---|---|---|---|---|---|---|
| 1 | 관계 | P_ MEMBER | MEMBER _NO | MEMBER | MEMBER_ NO | RI_0001 | 1:1 | |
| 2 | 관계 | PROJECT | PRJ_NO | PROJECT_ SPLY | PRJ_ NO | - | 1:N | |
| 4 | 코드 | CODE_ MASTER | CD | PROJECT | DEPT_ CD | CD_0002 | | TYPE_CD = '003' |
| 3 | 코드 | CODE_ MASTER | KIND_GB | PAPER | KIND_ GB | CD_0001 | | TYPE_CD = '004' |
| 5 | 관계 | PRODUCT | PRODUCT_ CD | PRODUCT_ DTL | PRODUCT_C D | - | 1:N | |

데이터베이스에 접속하여 DBMS에 부합된 테이블 및 컬럼의 레이아웃 정보를 추출한다. 컬럼 레이아웃은 각 DBMS의 종류에 따라 서로 상이하며 오라클의 경우 ALL_TABLES와 ALL_TAB_COLS 테이블을 조회하여 추출할 수 있다. MS-SQL의 경우 SYSOBJECTS와 SYSCOLUMNS를 조회하여 추출한다. 〈표 3-3〉은 오라클의 ALL_TAB_COLS 테이블을 조회하여 컬럼 레이아웃을 추출하는 예이다.

또한 일반적인 컬럼 레이아웃 외에도 관계 정의 등의 정보를 추출해야 하며, 이러한 정보는 제약조건 정보를 수록하고 있는 테이블을 통하여 추출할 수 있다. 〈표 3-4〉의 스크립트는 오라클에서 ALL_CONS_COLUMNS, ALL_CONSTRAINTS 테이블로부터 제약조건을 추출하는 스크립트이다. 키컬럼과 테이블, 외래키 컬럼과 테이블 외에 제약조건명・업데이트 조건・삭제 조건 등의 부가정보를 포함하여 추출한다.

〈표 3-3〉 테이블 및 컬럼 정보 추출 스크립트 예시

```
SELECT
  ROWNUM AS  "번호",
  T.TABLE_NAME AS "테이블명",
  T.COLUMN_NAME AS "컬럼명",
  T.COLUMN_ID AS "순서",
  T.DATA_TYPE AS "자료형",
  T.DATA_LENGTH AS "길이",
  T.DATA_PRECISION AS "정밀도",
  DECODE(T.NULLABLE, 'N' , 'NOT NULL' , 'Y', ' ') "널허용",
  T.DATA_DEFAULT AS "기본값",
  C.COMMENTS AS "코멘트"
FROM
  ALL_TAB_COLS T , ALL_COL_COMMENTS C
WHERE
  T.OWNER = C.OWNER AND
  T.TABLE_NAME = C.TABLE_NAME AND
  T.COLUMN_NAME = C.COLUMN_NAME AND
  T.OWNER = '[이용자]'
  ORDER BY T.TABLE_NAME, T.COLUMN_ID ASC ;
```

〈표 3-4〉 테이블 및 컬럼 관계 정보 추출 스크립트 예시

```
SELECT PK.OWNER AS PK_OWNER,
          PK.TABLE_NAME AS PKTABLE_NAME,
          PKACC.COLUMN_NAME AS PKOLUMN_NAME,
          FK.OWNER AS FK_OWNER,
          FK.TABLE_NAME AS FKTABLE_NAME,
          FKACC.COLUMN_NAME AS FKOLUMN_NAME,
          PKACC.POSITION AS KEY_SEQ,
          FK.DELETE_RULE AS DELETE_RULE,
          FK.CONSTRAINT_NAME AS FK_NAME,
          PK.CONSTRAINT_NAME AS PK_NAME,
          DECODE(PK.CONSTRAINT_TYPE, 'P', 'PRIMARY', 'UNIQUE') AS PK_UK
FROM  ALL_CONS_COLUMNS PKACC, ALL_CONS_COLUMNS FKACC,
       ALL_CONSTRAINTS   PK,     ALL_CONSTRAINTS    FK
WHERE PKACC.CONSTRAINT_NAME = PK.CONSTRAINT_NAME
 AND PKACC.OWNER = PK.OWNER
 AND PKACC.TABLE_NAME = PK.TABLE_NAME
 AND FKACC.CONSTRAINT_NAME = FK.CONSTRAINT_NAME
 AND FKACC.OWNER = FK.OWNER
 AND FKACC.TABLE_NAME = FK.TABLE_NAME
 AND PKACC.POSITION = FKACC.POSITION
 AND PK.CONSTRAINT_NAME = FK.R_CONSTRAINT_NAME
 AND PK.OWNER = FK.R_OWNER
 AND PK.CONSTRAINT_TYPE IN ('U', 'P')
 AND FK.CONSTRAINT_TYPE = 'R'
 AND PK.OWNER = '[이용자]' ;
```

## 1.2 메타데이터 분석

메타데이터 분석은 데이터 관리 문서와 실제 운영시스템간의 구조 정의를 비교하는 것으로 누락된 데이터 구조(구조적 완전성)와 불일치 유형(구조적 일관성)을 파악하는 절차이다.
사전에 취합된 테이블 및 컬럼, 관계 목록과 추출된 운영시스템의 테이블 및 컬럼, 관계 목록을 서로 대조하여 불일치 사항을 상세히 기재한다. 메타데이터의 불일치 사항 분석 시 엑셀의 비교 분석 함수를 사용하거나 별도의 데이터베이스에 업로드하여 분석 대상간 조인을 수행하여 불일치 내역을 추출한다.
메타데이터 분석을 통하여 수행할 수 있는 분석 유형은 테이블명 누락・불일치・컬럼 누락・컬럼명 불일치・자료형 불일치 등이 있다. 해당 분석을 실시하여 불일치한 테이블 및 컬럼, 관계 누락 사항을 기재하고 누락된 유형을 각각 명시한다.

〈표 3-5〉 테이블명 불일치 분석 예시

| DB명 | 문서 테이블명 | 운영 시스템 테이블 명 | 일치여부 |
|---|---|---|---|
| 통계 DB | TC_CD_INFO | TC_CD_INFO | 일치 |
| | … | … | … |
| | TH_ITM_PERIOD_HIST_OLD | 없음 | 누락 |
| | TN_ORG | TN_ORG | 일치 |
| | TN_PRS | TN_PERS | 불일치 |

〈표 3-6〉 컬럼명 및 자료형 불일치 분석 예시

| 물리 테이블 명 | 물리 컬럼명 | 문서 컬럼명 | 물리 자료형 | 문서 자료형 | 일치여부 |
|---|---|---|---|---|---|
| CHANGE_HIS | HIST_NO | HIST_NO | NUMBER | NUMBER | 일치 |
| DISTRICT_CODE | CLASSID | CLASS_ID | CHAR(10) | CHAR(10) | 컬럼 불일치 |
| … | … | … | … | … | … |
| CONN_INFO | INFO_TI | - | VARCHAR (128) | - | 컬럼 누락 |
| HOUSE_LEDGER | LDGR_TI | LDGR_TI | VARCHAR (128) | VARCHAR (50) | 자료형 불일치 |

불일치 구조 이외의 데이터 표준화와 관련된 추가 발견사항이 있으면 기재한다. 〈표 3-7〉은 데이터간의 비표준화 사항을 발견하여 목록으로 작성한 예이다.

〈표 3-7〉 비표준화 분석 예시

| 영문테이블명 | 한글 테이블명 | 영문컬럼명 | 한글컬럼명 | 자료형 | 길이 | 비 고 |
|---|---|---|---|---|---|---|
| BBS_CL01 | 고객문의 | READ_CNT | 조회수 | INTEGER | 11 | 조회수를 나타내는 컬럼인데 Data type이 INTEGER와 CHAR 두 종류가 사용됨 |
| BBS_CL02 | FAQ | READ_CNT | 조회수 | VARCHAR | 10 | |
| CUSTOMER | 고객 | CUSTOMER_ID | 고객ID | VARHAR | 20 | ID 컬럼의 Data type이 고객정보에서는 VARCHAR로 관리되며 주문정보에서는 CHAR로 관리됨 |
| ORDER | 주문 | CUST_ID | 고객ID | CHAR | 10 | |
| PAPER | 논문 | AUTHOR_ NAME | 저자명_한글 | VARCHAR | 128 | 각 컬럼명이 나타내는 의미의 정의가 일관되지 않게 정의됨(저자명을 NAME, NM, ENGNM, ENAME 등 다양하게 사용) |
| ARTICLE | 기사 | AUTHOR_NM | 저자명_한글 | VARCHAR | 128 | |
| | | AUTHOR_ ENGNM | 저자명_영문 | VARCHAR | 128 | |
| PUBLICATION | 간행물 | ENAME | 저자명_영문 | VARCHAR | 128 | |

## 2. 컬럼 속성 분석

컬럼 속성분석은 대상 컬럼의 총건수·유일값 수·NULL값 수·공백값 수·최대값·최소값·최대 빈도값·최소 빈도값·수치형의 경우 분산, 표준편차 평균의 기초 집계값 등을 각각 산출하여 그 값이 유효범위 내에 존재하는가 여부를 일차적으로 판단한다.

도메인의 값 범위·기본값·필수조건·유일성 판단 등의 기본적인 유효성 확인 절차를 육안검사와 관리문서 대조 등의 작업을 거쳐서 오류유형을 발견한다.

| No | 테이블명 | 컬럼명 | 총건수 | 유일값수 | NULL 개수 | 공백 개수 | 최대값 | 최소값 | 최대 빈도값 | 최대 빈도수 | 최소 빈도값 | 최소 빈도수 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CUSTOMER | U_ID | 24,351 | 24,351 | 0 | 0 | zzang01 | aeof032 | aeof032 | 1 | zzang01 | 1 |
| 2 | CUSTOMER | R_SN | 24,351 | 24,351 | 0 | 0 | 8510830-1****** | 500101-2****** | 510830-1******* | 1 | 811230-1******* | 1 |
| 3 | CUSTOMER | NAME | 24,351 | 19,013 | 0 | 5 | 홍길동 | Larry P English | 김선아 | 21 | 박정아 | 1 |
| 4 | CUSTOMER | REG_DT | 24,351 | 5,003 | 10 | 0 | 9999-12-31 | 1900-01-01 | 1999-01-01 | 532 | 2008-01-01 | 1 |
| 5 | CUSTOMER | ZIP_CD | 24,351 | 2,031 | 0 | 0 | 999-999 | 000-000 | 000-000 | 132 | 111-111 | 1 |
| 6 | CUSTOMER | EMAIL | 24,351 | 23,221 | 100 | 0 | zinks@zinks.com | abs@abc.com | dup@error.com | 3 | clean@clean.co.kr | 1 |
| 7 | CUSTOMER | TYPE | 24,351 | 12 | 0 | 0 | 12 | 1 | 10 | 1321 | 4 | 50 |
| : | : | : | : | : | : | : | : | : | : | : | : | : |
| 20 | APPLY | APP_NO | 130,001 | 130,001 | 0 | 0 | 141,020 | 1 | 141,020 | 1 | 1 | 1 |
| 21 | APPLY | U_ID | 130,001 | 20,880 | 0 | 0 | zzang01 | aeof032 | manyapp01 | 135 | minor03 | 1 |
| 22 | APPLY | STAT | 130,001 | 130,001 | 0 | 5 | 99 | 01 | 01 | | 99 | 1 |
| 23 | APPLY | REG_DATE | 130,001 | 130,001 | 0 | 0 | 2008-01-01 | 1900-01-01 | 1900-01-01 | 109 | 2008-01-01 | 1 |
| 24 | APPLY | CANCEL_TYPE | 130,001 | 10 | 128,980 | 0 | 09 | -1 | 01 | 132 | 06 | 10 |
| 25 | APPLY | CANCEL_DT | 130,001 | 988 | 129,031 | 0 | 2008-01-01 | 1999-01-31 | 2007-12-24 | 501 | 1999-01-01 | 1 |

〈그림 3-1〉 컬럼 프로파일링 분석 예시

〈그림 3-1〉은 컬럼 프로파일링 분석 결과를 표 형태로 집계한 결과이다. 위의 예제 파일은 SQL문을 활용하여 기본적으로 프로파일링에 필요한 자료를 취합하여 스프레드쉬트나·엑셀·CSV의 형태로 저장한 문서파일이다. 위의 컬럼 프로파

일링 분석 결과를 토대로 몇 가지 오류로 추정되는 사실을 발견할 수 있다.

① 고객의 성명(CUSTOMER.NAME)은 공백 값이 5건이 있다.
② 고객의 등록일(CUSTOMER.REG_DT)에 누락된 등록일이 10건 있으며, 일부 유효하지 않은 날짜 값(1900-01-01, 9999-12-31)들이 있다.
③ 고객의 우편번호코드(CUSTOMER.ZIP_CD)에는 일치하지 않는 값(999-999, 000-000)이 있다.
④ 고객의 Email(CUSTOMER.EMAIL)에는 누락된 값이 100건이 있으며, 중복된 email값인 dup@error.com 이 3건이 존재한다.
⑤ 주문신청의 취소유형(APPLY.CANCEL_TYPE)의 NULL 건수와 취소신청일(APPLY.CANCEL_DT)의 NULL 건수가 서로로 상이하다.
⑥ 주문신청의 취소유형(APPLY.CANCEL_TYPE)코드는 01에서 09까지의 코드외에 '-1'의 코드가 사용되었다.

①의 사례에서 유추되는 것은 마스터성 테이블의 자연식별자 컬럼은 기본적으로 NULL이 아니어야 하나 NULL값이 존재하는 완전성과 관련된 오류 사례이며, ②와 ③의 경우에는 유효성과 관련된 오류사례로 볼 수 있다. ④의 경우에는 유일성에 위배되는 오류로 볼 수 있으며, ⑤의 경우에는 복수 컬럼간 업무관계상의 오류로 볼 수 있다. ⑥의 경우에는 표준코드에 위반되는 코드 -1을 사용하는 코드 일관성 오류로 볼 수 있다.
앞의 사례는 직관적으로 오류의 가능성이 보이는 데이터를 지적한 사항이며 이는 오류일 가능성이 높은 오류 추정 데이

터이다. 이러한 오류 추정 데이터는 향후 프로파일링 결과 리뷰단계에서 비 유효 값으로 확인하기 위한 기초 데이터로 활용된다.

앞의 프로파일링에서는 단일 컬럼 값의 유효범위를 기준으로 직관적인 분석만이 가능하며 보다 상세한 데이터 유형을 발견하기 어렵다. 또한 날짜유형의 경우에는 상한, 하한의 경우에 해당되는 값만으로는 그 유효범위에 위배되는 데이터를 가려내기에는 어렵다. 날짜의 특성상 월별로 지정된 말일이 서로 상이하고(30일, 31일), 2월의 경우에는 해당 년도의 윤년 여부에 따라 월의 말일이(28일, 29일) 결정되기 때문이다.

그 외에도 주민등록번호・사업자등록번호・ISBN・ISSN 등의 번호 형식이 있는 특수 도메인에 해당되는 컬럼은 앞의 프로파일링 방법만으로는 그 진위 여부를 정확히 파악하기 힘들다. 이러한 데이터들의 오류현상을 발견하고 검증하기 위해서는 별도의 함수나 프로시저를 별도로 작성할 필요가 있다. 이러한 날짜유형・패턴・특수도메인 등은 다음 절에서 보다 상세하게 다룬다.

프로파일링 기법을 활용하여 컬럼 분석을 실시함에 앞서 수집되어야 할 기초 데이터가 필요하다. 이러한 데이터는 컬럼 프로파일링 분석을 위해 활용될 통계적 수치 데이터이며 컬럼의 총건수・NULL 값의 개수・공백과 0값의 개수・최대값・최소값・유일한 값의 개수・최대길이・최소길이・최대 빈도값・최소 빈도값 등이 해당된다.

실제 운영 중인 데이터베이스로부터 프로파일링 분석을 위한 기초 통계데이터를 SQL문을 활용하여 추출한다.

〈표 3-8〉의 SQL문을 진단 대상 컬럼의 형식에 맞추어 수행하고, 해당 출력 결과를 취합하여 저장한다. 아래 스크립트는 오라클을 기준으로 작성한 예제이며, 다른 DBMS에서는 문법

에 맞게 변형해야 한다.

〈표 3-8〉 분석 데이터 수집을 위한 SQL

① 전체레코드수, 최대값, 최소값, 최대길이, 최소길이 등

```
SELECT COUNT([컬럼]), MAX([컬럼]), Min([컬럼]),
       MAX(LENGTH([컬럼])), MIN(LENGTH([컬럼]))
FROM   [테이블] ;
```

② 유일한 값 개수

```
SELECT DISTINCT COUNT([컬럼])
FROM   [테이블] ;
```

③ NULL 값 개수

```
SELECT COUNT([컬럼])
FROM   [테이블]
WHERE [컬럼] IS NULL ;
```

④ 최대빈도값, 최소빈도값

```
SELECT [대상컬럼], CNT
FROM
        ( SELECT  [대상컬럼], COUNT([대상컬럼]) as CNT
          FROM   [대상테이블]
          GROUP BY [대상컬럼]
          ORDER BY CNT DESC      /* DESC:최대빈도, ASC:최소빈도 */ )
WHERE ROWNUM <= [개수] ;
* [개수]는 최대 빈도가 높은 것을 몇 개까지 추출할 것인가를 의미
```

※ MS-SQL의 경우에는 다음과 같이 사용한다.

```
SELECT TOP [개수] [컬럼], count(*) as CNT
FROM   [테이블]
GROUP BY [컬럼]
ORDER BY CNT DESC ;
```

특정 컬럼에 대하여 데이터의 레코드 전수를 육안으로 검사해야할 경우 해당 컬럼의 개별 값과 그 발생 빈도를 검사한다. 육안 검사는 사용된 값의 종류가 한정된 코드성 컬럼이나 구분·여부 컬럼 등의 검사에 활용된다. 유일한 키컬럼, 값의 종류 혹은 분포가 다양한 컬럼은 전수검사를 실시하지 않고 패턴분석을 활용한다. 〈표 3-9〉은 컬럼의 개별 값 및 빈도를 조회하는 SQL 문이다.

〈표 3-9〉 개별값 및 발생빈도를 조회하는 SQL

```
SELECT [대상컬럼], COUNT(*) as CNT
FROM   [대상테이블] SAMPLE ( [샘플링] )
GROUP BY [대상컬럼] ;
```

앞서 소개된 측정 SQL을 유형화하여 프로시저, 배치 스크립트 혹은 애플리케이션을 작성할 경우, 프로파일링 분석을 쉽게 할 수 있는 품질관리 도구를 제작할 수 있다.

## 3. 유형별 프로파일링 기법

### 3.1 누락 값 분석

누락 값 분석은 반드시 입력되어야 하는 데 값의 누락이 발생된 컬럼을 발견하는 절차이다. 누락 값 분석은 NULL값의 분포와 공백값(‘’), 숫자 ‘0’ 등의 분포를 파악하여 실시 한다.

| NO | 테이블명 | 컬럼명 | 컬럼 유형 | 컬럼 길이 | 기본값 | 총 건수 | NULL 건수 | 공백건수 |
|---|---|---|---|---|---|---|---|---|
| 1 | BOOK_MST | IDENTIFIER | CHAR | 20 | | 13,241 | 0 | 0 |
| 2 | BOOK_MST | TITLE | VARCHAR | 100 | | 13,241 | 10 | 13 |
| 3 | BOOK_MST | CLASSIFICATION | VARCHAR | 255 | | 13,241 | 0 | 57,332 |
| 4 | BOOK_MST | SIZE | NUMERIC | 10 | 0 | 13,241 | 0 | 0 |
| 5 | BOOK_MST | FORMAT | VARCHAR | 64 | | 13,241 | 0 | 12 |
| 6 | BOOK_MST | TYPE | VARCHAR | 10 | | 13,241 | 0 | 0 |
| 7 | BOOK_MST | KEYWORD | VARCHAR | 255 | | 13,241 | 6,589 | 0 |
| 8 | BOOK_MST | LANGUAGE | VARCHAR | 10 | | 13,241 | 0 | 12 |
| 9 | BOOK_MST | PAGE | INT | 10 | 0 | 13,241 | 0 | 0 |
| 10 | BOOK_MST | SEARCH_FLG | VARCHAR | 10 | | 13,241 | 13,241 | 0 |

〈그림 3-2〉 누락 값 분석 예시

도서 마스터 테이블(BOOK_MST)의 제목(TITLE)은 도서 테이블의 대표성이 있는 컬럼이나 NULL 값과 공백 값이 23건이 혼재되어 있다. 이는 모두 필수 컬럼의 누락과 관련된 완전성의 오류데이터로 볼 수 있다. 그 외에 자료형태(FORMAT), 언어(LANG) 등의 컬럼 또한 공백으로 채워진 완전성에 위배된 데이터로 볼 수 있다.

도서 마스터 테이블의(BOOK_MST)의 검색플래그(SEARCH_FLG) 컬럼은 총건수와 NULL 건수가 모두 동일한 컬럼이다. 이 컬럼은 데이터베이스에 설계되어 있지만 실제 사용되지 않는 미사용 컬럼으로 볼 수 있다.

## 3.2 값의 허용범위 분석

값의 허용 범위 분석은 컬럼의 속성 값이 가져야 할 범위 내에 속성 값이 있는지의 여부를 파악하는 것이며 이는 해당 속성의 도메인의 유형에 따라 그 범위가 결정된다. 측량의 단위·자료형의 크기·실수형의 경우 자릿수와 소수점 · 정밀

도 등이 주요 측정 대상이 된다.

| NO | 테이블명 | 컬럼명 | 컬럼 길이 | 컬럼 길이 | 총 건수 | NULL 건수 | 공백 건수 | 최대값 | 최소값 | 최대 길이 | 최소 길이 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | MAGAZINE | MAG_ID | CHAR | 10 | 131 | 0 | 0 | MG20071201 | MG19890101 | 10 | 10 |
| 2 | MAGAZINE | PUB_DT | VARCHAR | 10 | 131 | 131 | 0 | 2007/11/30 | 1989/01/01 | 10 | 10 |
| 3 | MAGAZINE | CLOSE_DT | VARCHAR | 10 | 131 | 13 | 0 | 2004/12/31 | | 10 | 0 |
| 4 | MAGAZINE | MAG_CD | VARCHAR | 4 | 131 | 0 | 0 | 0034 | 0001 | 4 | 4 |
| 5 | MAGAZINE | CYCLE_CD | VARCHAR | 2 | 131 | 0 | 0 | 09 | 01 | 2 | 2 |
| 6 | MAGAZINE | NO | NUMBER | 10 | 131 | 3 | 0 | 121 | -999 | 3 | 0 |
| 7 | MAGAZINE | VOL | NUMBER | 10 | 131 | 5 | 0 | 121 | -999 | 3 | 0 |
| 8 | MAGAZINE | TVOL | NUMBER | 10 | 131 | 5 | 0 | 100 | -999 | 3 | 0 |
| 9 | MAGAZINE | PAGES | NUMBER | 10 | 131 | 14 | 0 | 312 | -1 | 3 | 0 |

〈그림 3-3〉 값의 허용범위를 위반하는 컬럼

〈그림 3-3〉을 살펴보면 잡지의 호 · 권 · 통권 · 페이지는 0이상의 값을 가져야 하나 -999, -1 등의 유효범위에 어긋나는 값들이 존재한다.

이러한 값들은 값의 허용범위를 벗어나는 오류데이터이거나 애플리케이션에서 NULL값 대신 임의로 -999 등의 무의미한 값을 부여한 데이터일 가능성이 높다. 상기 값들은 업무 담당자와의 확인을 통해 진위 여부를 결정해야 된다. 별도의 의미를 가지지 않는 데이터로 확정될 경우, 상기 값들은 유효성 관점에서 모두 오류데이터로 판명되어 NULL 값으로 개선되어야 한다.

〈표 3-10〉 허용범위 위반 데이터 검증을 위한 SQL 예제

```
SELECT [컬럼], [키컬럼]
FROM   [테이블]
WHERE ([컬럼] < [시작범위]) OR
      ([컬럼] > [끝범위]) ;
```

## 3.3 허용 값 목록 분석

허용 값 목록 분석은 해당 컬럼의 허용값 목록이나 집합에 포함되지 않는 값을 발견하는 분석 방법이다. 허용값 목록 분석을 수행하기 위해서는 분석 대상 컬럼의 개별 값과 발생 빈도를 조사한다. 값의 유무나 여부를 나타내는 컬럼과 값이 명확히 정의되어 있는 유효 값의 컬럼, 표준화 되어 있지 않은 코드성 컬럼이 허용값 목록 분석 대상에 해당된다.

| CARD_TYPE | COUNT(CARD_TYPE) |
|---|---|
| A | 12,130 |
| B | 21,310 |
| C | 1,354 |
| D | 7,908 |
| 0 | 32 |
| 1 | 251 |
| NULL | 1,240 |

| USE_YN | COUNT(USE_YN) |
|---|---|
| Y | 42,130 |
| N | 61,211 |
| y | 123 |
| n | 68 |
| 0 | 5 |
| 1 | 11 |
| NULL | 1,023 |

〈그림 3-4〉 허용값 목록 분석

〈그림 3-4〉에서 CARD_TYPE은 'A', 'B', 'C', 'D'의 값을 사용해야 하나 실제 값에서는 '0', '1' 등의 값이 혼용되어 사용되었으며, 사용여부(USE_YN) 컬럼은 여부의 의미가 'Y', 'N'으로 구분되어 표기되어야 하나 '0', '1', 'y', 'n' 등의 의미가 유사하나 다른

값들이 혼용되어 사용되었다.

〈표 3-11〉 허용값 목록 위반 데이터 검증을 위한 SQL 예제

```
SELECT [컬럼], [키컬럼]
FROM   [테이블]
WHERE ([컬럼] NOT IN ( '[유효값1]', '[유효값2]', ..... , '[유효값N]' ) ;
```

## 3.4 문자열 패턴 분석

컬럼 패턴은 컬럼 속성 값의 특성을 문자열로 도식화한 것으로서 값의 특성이 문자열로 반복되고 변형되는 대표적인 모형을 미리 정형화하여 해당 컬럼의 특성을 파악하기 쉽게 해 놓은 데이터 표현 방법 중 하나이다.

일반적으로 키 컬럼이나 분포가 다양한 대용량 컬럼의 레코드의 모든 수를 육안 분석하는 것은 현실적으로 어려우므로 특정 번호(주민번호, 사업자등록번호), 코드성 컬럼 혹은 일부 문자로 정형화되어 발생 유형이 단일화 가능한 컬럼에 대하여 패턴 분석을 적용한다. 패턴 유형은 육안으로 식별할 수 있는 수준에서 패턴을 적용해야 하며 그 패턴의 종류가 너무 다양할 경우는 적용하지 않는다. 〈그림 3-5〉는 육안으로 보았을 경우 오류로 추정되는 데이터를 표시하였다.

| PAPER_NO | AUTHOR | LANG | ISSN | FILD_ID | FILE_PATH |
|---|---|---|---|---|---|
| 2007-1001 | 김논문 | KOR | 1225-6081 | KR010001 | /PUB/KR |
| 2007-1002 | 아시모 | JPN | 12257082 | JP010002 | /PUB/JP |
| 2007-1003 | E. A. JACK | ENG | 1225-608X | EN010003 | /PUB/EN |
| 2007-1004 | 박신양 | Kor | 1225-608 | KR010006 | C:₩PUB/KR |
| 2007-1005 | 아사코 | Japan | 1225-6085 | JP030002 | /PUB/JP |
| 2007-1006 | 김치국 | KOR | 1225-6086 | KR010005 | /PUB/KO |
| 2007-1007 | 스즈키 | JPN | 1225-6087 | KR010008 | C:₩PUB/JP |
| 2007-1008 | J.E. OLSON | ENG | 1225-6088 | KR080003 | /PUB/EN |
| 2007-1009- | 김정은 | KOR | 1225-6089 | KR080001 | /PUB/KR |

〈그림 3-5〉 육안 분석을 통해 발견된 오류 데이터

제공언어(LANG)와 같은 컬럼의 경우에는 개별 값 및 빈도를 전수검사를 하여도 그 발견되는 유형이 다양하지 않으므로 분석이 어렵지 않으나, 논문번호(PAPER_NO)의 경우는 주요키(Primary Key) 컬럼이므로 전수의 오류 유형을 육안으로 분석하는 것은 어렵다. 또한 해당 테이블의 레코드수가 매우 많다면 각 컬럼을 육안검사로 오류유형을 발견하기는 거의 불가능하다. 따라서 이러한 컬럼들은 특정 패턴을 발견하는 함수와 문자열의 길이를 반환하는 함수를 작성하여 분석하는 것이 바람직하다.

패턴 발견 함수는 보통 전수의 레코드를 검사하고, 컬럼의 스트링을 처음부터 끝까지 조사하기 때문에, 그 수행 작업이 운영시스템에 부하를 줄 수 있으므로 수행시간의 스케줄링과 샘플링을 고려해야 한다. 〈표 3-12〉는 샘플링의 비율을 50%로 적용하여 논문번호의 발생 패턴과 개별 건수를 출력하는 예제이다.

아래 예제의 GET_PATTERN(문자열) 함수는 문자열의 값을 파라미터로 입력받아 입력 문자가 영문자일 경우 C, 숫자일 경우 9, 공백일 경우 S를 반환하는 함수이다. 패턴발견 함수는

문자열을 입력받아서 문자열의 시작부터 끝까지 매칭하고, 문자열의 아스키코드와 대치문자의 아스키코드를 비교하여 해당되는 문자열을 원하는 패턴문자로 대치하여 그 결과 값을 반환한다.

〈표 3-12〉 패턴별 개별값 목록 나열 SQL 예제

```
SELECT GET_PATTERN(PAPER_NO) as PAT, COUNT(*) as CNT
FROM   PAPER SAMPLE(50)
GROUP BY GET_PATTERN(PAPER_NO) ;
```

| ISSN | CNT |
|---|---|
| 9999-9999 | 13,252 |
| 999-99999 | 1 |
| 9999-999C | 3 |
| 9999-9SSS | 5 |

| FILE_PATH | CNT |
|---|---|
| /CCC/CC | 6,098 |
| /CCC/CCC | 7,139 |
| C:₩CCC/CC | 10 |
| D:₩CCC/CCC | 14 |

〈그림 3-6〉 패턴별 개별 값 목록

〈그림 3-6〉은 패턴함수를 ISSN과 FILE_PATH에 적용하였을 경우 발견되는 패턴의 예이다. ISSN번호는 앞의 네 자리는 숫자, 중간에 하이픈, 뒤의 네 자리는 숫자의 형태로 이루어져야 하나 패턴 규칙에 위배되는 사항이 9건이 검출된다. 또한 FILE_PATH에서도 비정상적인 패턴 값이 24건이 검출된다.

## 3.5 날짜유형 분석

일반적으로 날짜유형을 표현할 경우 다음의 두 가지의 방법을 많이 사용한다. 첫째 DBMS에서 제공하는 DATETIME의 유형을 사용하는 경우이며, 둘째는 문자형에 날짜패턴을 적용하여 활용하는 경우이다. 전자의 경우에는 날짜유형의 유효범위 검사는 불필요하다. DATETIME 유형은 DBMS에서 제공하는 TIMESTAMP를 기준으로 날짜의 의미와 년·월·일·요일·시간 정보까지 매우 정확하게 제공되기 때문이다. 따라서 날짜 및 시간 형식 데이터는 DATETIME 유형을 사용하는 것이 바람직하나 비즈니스의 요구사항에 따라 문자형으로 대체하여 사용하기도 한다.

DATETIME 유형은 데이터베이스관리시스템의 특성에 따라서 적용 가능한 년도가 차이가 있으며(특정 시스템은 1000년도 이전의 날짜를 지원 못하기도 함), 고문서관리 시스템이나 역사정보시스템 등과 같이 오래된 과거 데이터를 다루거나 기원전 년도 등의 날짜를 다루는 시스템의 경우에는 DATETIME의 유형으로는 그 날짜와 시간을 정확히 표현하기 쉽지 않으므로 후자의 형태와 같이 텍스트 형태의 자료형에 날짜유형의 패턴을 통일시켜 활용한다. 날짜의 유효성과 관련된 문제는 후자의 경우 빈번히 발생한다.

〈그림 3-7〉에서 보면 시작일과 종료일의 날짜유형은 YYYY-MM-DD의 형태를 전반적으로 사용하는 것을 알 수 있다.

①은 유효범위에 어긋나는 데이터의 범위 값이 사용되었다. 다만 종료일이 너무 길기 때문에 최대 날짜에 속하는 값으로 9999-12-31 값을 임의로 입력한 것으로 보인다. 이러한 날짜 데이터의 유효 여부는 업무 담당자와의 협의를 통하여 결

| PRJ_NO | PRJ_NM | DUTY_NM | START_DT | END_DT |
|---|---|---|---|---|
| 1998-1001 | Data Quality Assesment Project | J. S. Suh | 1998-01-03 | 9999-12-31 ① |
| 2003-1002 | Data Quality Dimensions | S..s. Shin | 2003-01-02 | 2004-12-31 |
| 2005-1003 | Real time Data Base | I. ch. Kim ② | 2005-02-31 | 2005-12-31 |
| 2005-1004 | Maintaining Data Quality during Ent.. | S. Y. Kim ③ | 20050329 | 2005-12-31 |
| 2006-1005 | Boosting Database Quality in .. | Park Jang | 2006-12-31 | 2006-10-30 |
| 2007-1006 | Quality Assurance in Database | P. Konglish ④ | 2007/04/31 | 2007-12-31 |

〈그림 3-7〉 날짜유형 분석 예시

정해야 할 사항이다.

②의 경우는 2월은 28일 또는 29일까지 있으나 31일이 입력이 되어 있다. 보통 말일을 나타내기 위해서 임의로 해당월에 31일 값을 입력한 사례로 보인다.

③의 경우는 YYYY-MM-DD 유형이 아닌 YYYYMMDD의 패턴으로 '-'이 빠진 형태로 일관성이 없이 입력된 형태로 보인다.

④의 경우는 4월은 30일까지만 있으나 이 또한 ②의 경우와 마찬가지로 말일을 표현하기 위해 임의로 31을 입력한 것으로 보인다.

①과 같은 예에서는 담당자가 값을 모르는 경우이거나, 값이 지정되지 않은 경우일 수 있다. 1998년도에 시작된 프로젝트가 계속사업이고 언제 끝날지 모르기 때문에 9999-12-31과 같은 값을 입력하는 현상이 실 사례에서는 빈번하게 발생된다.

종료일은 실제 종료되었거나 종료될 날짜 값을 의미하므로 종료시점을 정확히 입력해야 하나 아직 종료되지 않은 사업을 언제 종료될지 모르는 막연한 시점으로 재해석해 아주 먼 미래의 시점을 입력한 것이다. 이러한 경우에는 막연한 시점의 값을 임의로 입력하는 것 보다 NULL값을 입력하고 사업 구분을 '계속사업'으로 의미를 명확히 하는 것이 바람직하다.

### 1) 날짜유형 발견

이러한 날짜유형을 발견하기 위해서는 별도의 날짜의 유효범위를 체크하는 함수를 작성하여 날짜 규칙에 어긋나는 데이터를 검증해야 한다. 날짜 유효성의 발견은 문자열의 패턴 발견과 유사하므로 문자패턴을 날짜패턴에 적용하여 활용하거나 날짜유형의 특성에 맞게 검증함수를 작성하여 분석한다.

〈표 3-13〉 날짜 패턴 발견 SQL

```
SELECT GET_PATTERN(START_DT) as START_DT, COUNT(*) as CNT
FROM   PROJECT
GROUP BY GET_PATTERN(START_DT) ;
```

| START_DT | CNT |
|---|---|
| 9999-99-99 | 101,131 |
| 9999/99/99 | 11 |
| 99999999 | 2 |
| 99-99-99 | 1 |

〈그림 3-8〉 날짜 패턴 개별 값 목록

〈그림 3-8〉은 시작일에 패턴 분석을 적용한 결과이다. 위의 예에서는 '9999-99-99'를 유효한 날짜 유형으로 볼 수 있으며 날짜에 '-' 이외의 '/' 등이 삽입되거나 구분자가 없는 숫자의 나열로 입력된 숫자 혹은 년도를 지키지 않은 비유효한 날짜 값이 총 14건이 발견되었다.

〈표 3-14〉 날짜 유형 검증함수 예시

```
  CREATE OR REPLACE FUNCTION CHK_DATE( P1 VARCHAR2, P2 VARCHAR2 )
RETURN VARCHAR2 IS
      V_TEST_DATE DATE;
  BEGIN
      IF P1 IS NULL THEN
          RETURN NULL;
      ELSE
          V_TEST_DATE := TO_DATE(P1,P2,'NLS_DATE_LANGUAGE = KOREAN');
          RETURN 'T';
      END IF;

      EXCEPTION
            WHEN OTHERS THEN
                RETURN 'F';
  END;
```

※ 본 함수는 오라클을 기준으로 작성되었으며, P1은 점검컬럼, P2는 점검컬럼의 날짜유형이 입력되며, 날짜값이 유효할 경우에는 'T', 비유효시에는 'F'의 값을 반환한다.
두 번째 파라미터는 TO_DATE함수의 FORMAT_MASK의 활용 예를 참조하기 바란다.

※ 다음은 날짜 유형을 검증하는 예제이다.

```
/* 년도-월-일 형식의 날짜 유형의 검증 */
SELECT [컬럼], [키컬럼]
FROM   [테이블]
WHERE CHK_DATE( [컬럼], 'YYYY-MM-DD' ) <> 'T' ;

/* 년도-월-일-24시간-분-초 형식의 날짜 유형의 검증 */
SELECT [컬럼], [키컬럼]
FROM   [테이블]
WHERE CHK_DATE( [컬럼], 'YYYY-MM-DD HH24:MI:SS' ) <> 'T' ;
```

※ ORACLE의 TO_DATE 함수는 년도 YYYY일 경우 자리수가 1자리 이상 4자리 미만의 년도 숫자를 모두 허용하므로, 자리수가 일치해야 하는 경우는 자리수 검증로직 점검 후 날짜 검증 함수를 적용하거나 패턴을 적용한 후 날짜를 검증해야 한다.

2) 날짜유형의 유효성 검증 SQL

날짜유형의 검증 방법은 날짜의 유효 패턴 및 길이를 검증한 후 CHK_DATE란 함수를 사용하여 그 날짜의 유효여부를 검증한다.

〈표 3-15〉 날짜유형 검증 SQL 활용 예시

```
SELECT [컬럼],[키컬럼]
FROM   [테이블]
WHERE LENGTH( REPLACE( [컬럼], '-', '' ) ) <> 6
   AND CHK_DATE( [컬럼], 'YYYY-MM-DD' ) <> 'T'
   AND [컬럼] IS NOT NULL ;
```

## 3.6 기타 특수 도메인(특정 번호 유형)의 분석

날짜 도메인과 유사하게 값의 패턴과 유효범위로 컬럼 값의 유효성을 확인하기에 보다 복잡한 구조를 가지는 번호체계가 존재한다. 주민등록번호・사업자등록번호・ISBN・ISSN 등의 특정 번호 형식이 있는 특수 도메인 컬럼의 경우가 이에 해당된다.

예를 들어 사업자등록번호는 123-45-67890와 같은 형식을 갖고 있다. 사업자 등록번호는 숫자 10자리와 2개의 '-'으로 구성되는데 처음 3자리는 청서 코드, 두 번째 2자리는 사업자 구분, 셋째 5자리는 일련번호를 의미한다.

사업자 등록번호는 별도의 검증로직이 있으며, 검증로직은 각각 자리수의 구성과 그 개별 값의 연산 결과와 관련되어 있다. 이는 부적절한 사업자 등록 번호의 사용을 막기 위하여 활용된다.

이러한 특수 도메인의 경우에는 도메인의 생성패턴 외에도 별도의 복잡한 검증을 위하여 보다 복잡한 방법이 요구되기도 한다. 일반적으로 번호 또는 코드의 검증을 위하여, 별도의 함수나 프로시저 등을 작성해야 한다. 〈표 3-16〉은 사업자 등록번호 검증을 위한 함수 작성의 예를 나타내고 있다.

〈표 3-16〉 사업자 등록번호 검증 함수의 예시

```
CREATE OR REPLACE FUNCTION CHK_WORK_NUM ( V_P1 IN VARCHAR2 )
RETURN VARCHAR2 IS
    V_TEST     VARCHAR2(20) ;
    CHK_SUM   VARCHAR2(20);
    SUM_MOD   NUMBER;
BEGIN
        IF V_P1 IS NULL THEN
            RETURN NULL;
        ELSE
            V_TEST := REPLACE(V_P1, '-','');
            IF LENGTH(V_TEST) <> 10 THEN
                RETURN 'F' ;
            END IF;
        END IF;
        SUM_MOD := 0;
        SUM_MOD := SUM_MOD + TO_NUMBER(SUBSTR(V_TEST,1,1)) ;
        SUM_MOD := SUM_MOD + MOD(TO_NUMBER(SUBSTR(V_TEST,2,1))*3,10)  ;
        SUM_MOD := SUM_MOD + MOD(TO_NUMBER(SUBSTR(V_TEST,3,1))*7,10)  ;
        SUM_MOD := SUM_MOD + MOD(TO_NUMBER(SUBSTR(V_TEST,4,1))*1,10)  ;
                                        …
                                    = 중략 =
                                        …
        SUM_MOD := SUM_MOD + MOD(TO_NUMBER(SUBSTR(V_TEST,8,1))*3,10)  ;
        SUM_MOD := SUM_MOD + FLOOR(TO_NUMBER(SUBSTR(V_TEST,9,1))* 5/10)  ;
        SUM_MOD := SUM_MOD + MOD(TO_NUMBER(SUBSTR(V_TEST,9,1))*5,10)  ;
        SUM_MOD := SUM_MOD + MOD(TO_NUMBER(SUBSTR(V_TEST,10,1)),10);
        IF MOD(SUM_MOD ,10) = 0 THEN
            RETURN 'T';
        ELSE
            RETURN 'F';
        END IF ;

     EXCEPTION
        WHEN OTHERS THEN
            RETURN '[ERROR]';
END;
```

특수 도메인은 해당되는 컬럼의 패턴분석을 통하여 일차적인 비유효 패턴을 발견하고 오류데이터의 검증은 컬럼 속성 값의 특수 도메인 검증 로직을 통하여 이루어진다.

## 3.7 유일값 분석

유일값 분석은 업무적 의미에서 유일해야 하는 컬럼에 중복이 발생되었는가를 파악하기 위한 것이다. 테이블의 식별자로 활용되는 컬럼 속성 값들이 주요 유일값 분석 대상이다. 예를 들면 고객 마스터 테이블의 주민등록번호, 사업자 마스터 테이블의 사업자등록번호 등이 이에 해당된다.

| No | 테이블명 | 컬럼명 | 총건수 | 유일값수 | NULL 개수 | 공백 개수 | 최대 빈도값 | 최대 빈도수 | 최소 빈도값 | 최소 빈도수 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CUSTOMER | U_ID | 24,351 | 24,351 | 0 | 0 | aeof032 | 1 | zzang01 | 1 |
| 2 | CUSTOMER | R_SN | 24,351 | 24,351 | 0 | 0 | 510830-1234567 | 1 | 811230-1234567 | 1 |
| 3 | CUSTOMER | NAME | 24,351 | 19,013 | 0 | 5 | 김선아 | 21 | 박정아 | 1 |
| 4 | CUSTOMER | REG_DT | 24,351 | 5,003 | 10 | 0 | 1999-01-01 | 532 | 2008-01-01 | 1 |
| 5 | CUSTOMER | ZIP_CD | 24,351 | 2,031 | 0 | 0 | 000-000 | 132 | 111-111 | 1 |
| 6 | CUSTOMER | EMAIL | 24,351 | 23,221 | 100 | 0 | dup@error.com | 3 | clean@clean.co.kr | 1 |
| 7 | CUSTOMER | TYPE | 24,351 | 12 | 0 | 0 | 10 | 1321 | 4 | 50 |

〈그림 3-9〉 유일값 분석 예시

〈그림 3-9〉에서 고객의 이메일은 유일해야 하나 중복된 이메일이 발견된다. 중복된 데이터는 컬럼의 총 건수에서 유일한 값의 개수를 뺀 값 값만큼 중복되었다고 볼 수 있다. 위의 예에서 최대 빈도값과 최대 빈도수를 확인해 보면 가장 많이 중복된 데이터를 확인 할 수 있다.

데이터의 중복이 발생될 경우 데이터의 신뢰성이 낮아지며 키 컬럼과 같은 식별자 컬럼의 값이 중복될 경우 해당 컬럼과 관계에 있는 모든 테이블에서 중복되는 수량만큼의 오류 데이터가 발생할 가능성이 높아지게 된다.

〈표 3-17〉 유일성 검증 SQL 예시

```
SELECT A.[컬럼], A.[키컬럼]
FROM   [테이블] A,
       ( SELECT B.[컬럼]
         FROM   [테이블] B
         GROUP BY B.[컬럼]
         HAVING COUNT(*) > 1 ) C
WHERE A.[컬럼] = C.[컬럼] ;
```

## 3.8 구조 분석

데이터 구조 분석은 구조 결함으로 인한 일관되지 못한 데이터를 발견하는 분석 기법으로, 관계분석・참조 무결성 분석・구조 무결성 분석 등으로 불리기도 한다. 즉, 데이터 구조 분석은 잘못된 데이터 구조로 인해 데이터 값에서 일관되지 못하거나 부정확한 값이 발견되는 현상을 파악하는 작업이다. 또한 데이터의 구조적 완전성 문제로 인해 데이터의 일관성이 결여되는 데이터 값을 발견하고, 사전에 정의된 구조 외에 누락된 구조를 발견하고 테스트하여 정확한 구조를 파악하는 것을 주된 목적으로 한다.

〈그림 3-10〉은 고객이 제품을 주문하는 ERD의 예이다. 고객과 주문은 1:M의 관계가 있으며 제품과 주문은 1:M의 관계에 있다. 이러한 논리 관계는 물리 관계로 변경 시 테이블과 테이블의 관계로 변환이 되며, 〈그림 3-11〉에서는 물리적 테이

블로 변환된 과정을 표현하고 있다.


〈그림 3-10〉 고객의 제품 주문 ERD 예시

고객의 고객ID(CUSTOMER.CUST_ID)와 주문의 고객ID(ORDER.CUST_ID)는 서로 주키와 외래키로 관계를 맺고 있으며, 제품의 제품코드(PRODUCT.PROD_ID)와 주문의 제품코드(ORDER.PRO_ID) 또한 서로 주키와 외래키로 관계를 맺고 있다.

TABLE_NAME : CUSTOMER

| CUST_ID | CUST_NM | CUST_ADR1 | CUST_TEL |
|---|---|---|---|
| JJS01 | 정재성 | 인천 부평구 청천동 | 111-2222 |
| PJS02 | 박종식 | 서울 강동구 고덕1동 | 111-1111 |
| KST01 | 김수택 | 서울 송파구 신천동 | 222-2222 |
| KCJ08 | 김창진 | 서울 구로구 구로5동 | 333-3333 |

TABLE_NAME : ORDER

| CUST_ID | PROD_ID | ORDER_DT | ORDER_QTY | PRICE |
|---|---|---|---|---|
| JJS01 | C0012 | 2008/01/31 | 2 | 1460000 |
| JJS01 | C0020 | 2008/02/21 | 1 | 180000 |
| KST01 | C0020 | 2007/12/05 | 1 | 180000 |
| KST01 | C0051 | 2008/01/12 | 2 | 420000 |
| PMS11 | C0061 | 2008/01/15 | 1 | 134000 |
| KCJ08 | C0061 | 2008/02/13 | 1 | 134000 |

TABLE_NAME : PRODUCT

| PROD_ID | PROD_NM | PRICE | VENDOR |
|---|---|---|---|
| C0012 | 컴퓨터 | 730000 | 삼성 |
| C0020 | 모니터 | 180000 | LG |
| C0051 | XBOX | 210000 | MS |
| C0061 | 프린터 | 134000 | HP |

〈그림 3-11〉 고객, 제품, 주문 테이블 예시

주문테이블(ORDER)에는 고객테이블에 존재하지 않는 고객의 ID가 잘못 사용된 예를 볼 수 있다. 해당 고객의 정보(고객ID가 PMS11)는 삭제하였지만 주문 내역에는 삭제하지 않아서 주문 내역이 남아 있는 경우이다. 이러한 경우에는 데이터의 일관성이 결여된 오류데이터로 볼 수 있다.

실무에서는 관계 제약조건을 DBMS에 직접 적용하지 않고 구조 변경의 유연성을 위해 방치하거나 응용프로그램에서 처리하는 사례가 빈번히 발생된다. 이렇게 관계 제약 조건을 응용프로그램에 누락된 상태로 방치하여 운영할 경우 일관되지 못한 데이터가 지속적으로 누적된다.

데이터 구조의 누락으로 인한 오류 데이터의 발생을 미연에 방지하기 위해서는 해당 데이터의 입력・수정・삭제 시 연관된 데이터의 처리 방법을 사전에 명확히 결정하고, 물리모델링에서 확정된 구조정보를 운영 데이터베이스에 참조 무결성 제약조건으로 적용해야 한다.

### 3.8.1 추가 구조의 발견

추가 구조의 발견 절차는 메타데이터 수집 단계에서 사전 정의한 관계 목록 이외에 누락된 관계 목록이 존재하는가를 파악하는 절차이다. 이 절차에서는 주요키・외래키・동의어 컬럼 등을 규명하고, 메타데이터 분석 단계에서 규명된 구조 정보 이외의 누락되었거나, 응용프로그램에서 처리되는 미 설계된 추가 구조 규칙을 발견하는 것을 주된 목적으로 한다.

누락된 추가 구조 발견을 위한 사전 파악할 정보는 다음과 같다.

- 테이블 및 컬럼 명칭의 영문화 및 약어사용 : 메타데이터의 의미

• NULL 및 유일성 제약 조건 : 키후보
• 주요키와 외래키 관계 : 테이블간 관계
• 중복 테이블 및 컬럼 : 동의어 컬럼 관계

사전 수집된 메타데이터를 토대로 테이블 및 컬럼의 영문명과 약어사용 분포를 파악하여 해당 컬럼의 정확한 의미와 역할을 파악한다. 또한 NULL 및 유일성 제약조건을 규명하여 주요키 후보를 확정한다. 확정된 주요키를 기준으로 외래키·동의어 컬럼·종속성 후보를 파악하여 관계 후보 목록을 작성한 후, 해당 컬럼간 포함률과 역포함률을 측정하여 관계를 규명한다.

### 1) 주요키

주요 키 컬럼은 일반적으로 완전성과 유일성을 동시에 만족해야 하는 특성이 있다. 또한 이미 컬럼 분석에서 각 컬럼의 NULL여부와 유일성 규칙을 규명하였으므로 그 특성을 활용하여 주요키 컬럼을 규명할 수 있다.

키 후보는 식별자로 활용되므로 그 메타데이터의 명칭이 식별자 형태인 것들이 해당된다. ORDER_NO, PRODUCT_ID, TASK_NUMBER, EQUIP_CD 등의 컬럼 명칭의 특성을 활용하여 육안으로 키 컬럼의 후보를 추정할 수 있다. 또한 컬럼의 명칭 이외의 테이블 및 컬럼 정의서와 용어 사전 등을 분석하여 각 컬럼이 어떠한 사항을 의미하는 가를 파악한다.

일반적으로 한 테이블당 적어도 하나 이상의 키 컬럼이 존재한다. 단, 자연키를 복합키로 사용하거나 다수의 키의 조합으로 키컬럼을 사용할 때의 예외사항이 있다. 하나의 테이블당 한 개 이상의 키 컬럼을 발견하지 못한 경우에는 자연키나 복합키를 사용했을 가능성이 있다.

### 2) 외래키

외래키 컬럼은 다른 테이블의 식별자와 관계에 있는 종속되는 테이블의 컬럼이다. 외래키 컬럼의 규명을 위해서는 우선 마스터성 테이블이나 부모테이블의 식별자를 도출하고 이 식별자에 해당되는 동의어 목록을 발견한다.
동의어 목록은 수집된 메타데이터와 컬럼 명칭을 토대로 발견하되 해당 동의어 컬럼의 업무 의미를 자세히 파악하기 위해, 육안분석이나 키 컬럼과 해당 동의어 컬럼을 매칭하여 분석한다.
외래키 컬럼의 발견은 단계적으로 수행한다. 주요 마스터성 테이블의 키 컬럼으로부터 계층적으로 관계에 있는 자식테이블의 동의어 컬럼을 식별하고, 자식테이블의 키 컬럼으로부터 연관된 손자테이블의 동의어 컬럼을 찾아가는 하향식으로 수행한다.

### 3) 동의어 목록

동의어 목록은 데이터의 구조를 발견하는데 매우 유용하다. 동의어는 앞서 언급한 관계를 규명하기 위한 키컬럼 동의어 외에도 중복 데이터 동의어 · 도메인 동의어 · 병합 동의어 등이 있다.
중복 데이터 동의어는 데이터 모델링 시 성능 개선 및 관리 효율의 목적으로 반정규화 과정을 거칠 때 사용되는 컬럼으로 동일 의미를 가지는 컬럼이 여러 테이블에 복사되어 활용되는 컬럼이다. 도메인 동의어는 날짜유형 · 도시명 · 우편번호 등의 동일한 도메인의 동일한 의미를 내포하는 동의어이다. 병합 동의어는 데이터 통합 시 활용되는 동의어로 분산시스

템에서 활용하는 컬럼을 하나의 통합시스템에서 활용할 때 사용되는 동의어이다.

이러한 동의어들 중에서 활용도가 높거나 데이터의 일관성이 중요한 컬럼을 후보로 도출한다. 후보로 도출된 컬럼은 기준 컬럼과 대상 컬럼의 포함률이나 중첩도를 측정하여 각 동의어 간의 관계를 규명한다.

〈표 3-18〉는 동의어 후보들 중에서 포함률을 측정하여 외래키 후보를 선정하는 예를 나타내고 있다. 포함률은 해당 주키 컬럼(기준 컬럼)과 외래키 컬럼(측정 대상 컬럼) 후보들 간 상호 일치 건수를 측정 대상 테이블의 레코드 건수로 나누어 백분율로 나타낸 것이며 역포함률은 주키와 외래키를 역으로 적용하여 측정한 것이다.

〈표 3-18〉 관계 포함률 측정 예시

| NO | 주키 테이블명 | 주키 컬럼 | 외래키 테이블명 | 외래키 컬럼명 | 포함율 | 역포함율 | 선정 | 비고 |
|---|---|---|---|---|---|---|---|---|
| 1 | T_MEM_B | MEMBER_NO | T_BIZ | MEMBER_NO | 99% | 60% | √ | |
| 2 | T_MEM_B | MEMBER_NO | T_BOARD | MEMBER_NUMBER | 100% | 30% | √ | |
| 3 | T_BOOTH | BOOTH_NO | T_BOARD | BOOTH_NUMBER | 20% | 30% | | |
| 4 | T_COM_DIV | DIV_NO | T_BOOTH | DIV_CD | 100% | 71% | √ | |
| 5 | T_BOOTH_B | BOOTH_BASE_NO | T_BOOTH | BOOTH_BASE_NO | 99% | 40% | √ | |
| 6 | T_COM_DIV | DIV_NO | T_BUYER | DIV_NUMBER | 98% | 54% | √ | |
| 7 | T_BOOTH | BOOTH_NO | T_CARD_C | P_BOOTH_CD | 12% | 14% | | |
| 8 | T_COM_MEM_TEMP | CMEMBER_NO | T_CARD_C | CMEMBER_NO | 100% | 60% | √ | |

### 3.8.2 구조 발견 절차

구조 분석은 누락값 · 유일값 · 유효값 분석 등이 완료된 시점 이후에 수행하는 것이 바람직하다. 사전에 수행된 프로파일링 분석에서 컬럼의 누락 값과 유일 값 특성이 파악된 후 이 자료를 토대로 데이터 구조를 파악하는 것이 용이하기 때문이다. 구조 발견 절차는 다음과 같다.

1) 마스터성 테이블이나 코드테이블과 같은 업무상 핵심 테이블을 규명한다.
2) 핵심 테이블의 주요키를 규명하고 관련 테이블 · 외래키 · 동의어 컬럼을 규명한다.
3) 주요키와 외래키의 관계 목록을 정리한다.
4) 하위 연관 관계있는 테이블을 규명하고 규명된 테이블을 기준으로 동의어 컬럼, 종속관계에 있는 컬럼을 도출하여 발견 과정을 확장한다.
5) 모든 테이블을 분석할 때까지 발견 절차를 반복한다.

### 3.8.3 데이터 구조 무결성 검증

사전 수집된 관계 목록과 새로 발견된 관계 목록을 취합하여 데이터 구조 무결성 위반 내역을 추출한다. 구조 분석을 수행하는 스크립트의 예는 〈표 3-19〉와 같다. 관계 분석의 데이터 검증을 위한 SQL문은 수행 속도의 문제 때문에 NOT IN이란 집합연산을 하지 않고, OUTER JOIN을 활용해 측정한다. 참조관계의 오류 목록에는 대상 컬럼 외에도 대상 테이블의 키컬럼 또한 명시하도록 한다.

〈표 3-19〉 구조 무결성 위반 데이터 검증 SQL 예시

```
SELECT ORDER1.CUSTOMER_ID, ORDER1.ORDER_ID
        FROM ORDER1 LEFT OUTER JOIN CUSTOMER ON
CUSTOMER.CUSTOMER_ID = ORDER1.CUSTOMER_ID
WHERE
CUSTOMER.CUSTOMER_ID IS NULL ;
```

〈표 3-20〉 구조 무결성 위반 내역 목록

| NO | 기준 테이블명 | 기준 컬럼명 | 대상 테이블명 | 대상 컬럼명 | 대상 키컬럼명 | 관계명 | 기수성 | 진단 건수 | 오류 건수 | 오류율 (%) | 비고 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | MEMBER | MEMBER_NO | MEMBER_DTL | MEMBER_NO | MEMBER_NO | RI_0001 | 1:1 | 9,084 | 121 | 1.33 | |
| 2 | PROJECT | PRJ_NO | PROJECT_SPLY | PRJ_NO | SPLY_ID | - | 1:N | 23,351 | 51 | 0.22 | |
| 3 | PRODUCT | PRODUCT_CD | REPAIR_DTL | PRODUCT_CD | REPAIR_SN | - | 1:N | 123,983 | 53 | 0.04 | |
| 4 | CUSTOMER | CUSTOMER_ID | ORDER | CUSTOMER_ID | ORDER_ID | RI_0003 | 1:N | 203,391 | 43 | 0.02 | |

## 4. 프로파일링 결과 리뷰 및 종합

관계 목록 이외의 통합코드 테이블을 참조하는 코드성 컬럼을 분석하는 스크립트는 〈표 3-21〉와 같다. 통합 코드는 그 활용 목적과 방법에 따라 별도의 구분코드나 추가 컬럼이 활용될 수 있으므로, 이와 관련된 제약조건을 검증 SQL문에 추가하여 활용한다.

〈표 3-21〉 코드 일관성 위반 데이터 검증 SQL 예시

```
SELECT PROJECT.DEPT CD,
       PROJECT.CD
FROM   PROJECT LEFT OUTER JOIN CODE_MASTER ON
CODE_MASTER.CD = PROJECT.DEPT_CD
AND    CODE_MASTER.TYPE_CD = '003' -- 유형코드 '003'을 사용
WHERE  CODE_MASTER.CD IS NULL
AND    PROJECT.DEPT_CD IS NOT NULL ;  -- DEPT_CD가 NULL값을 허용
```

프로파일링 분석이 완료되면 프로파일링 결과물을 취합하여 관련 업무 담당자에게 전달하여 오류데이터 발생 컬럼의 진위 확인을 요청한다. 오류확인 여부와 함께 담당자의 검토 의견을 상세히 작성하도록 요청한다.

## 4.1 필수 값이 누락된 컬럼

필수항목으로 입력되어야 할 항목이나 NULL 또는 ''(공백)이 입력된 컬럼을 모두 나열하여 목록을 작성한다.

〈표 3-22〉 누락된 값 목록 예시

| 테이블명 | 컬럼명 | 데이터 타입 | 길이 | 기본 값 | 총건수 | 공백 수 | NULL 건수 | 오류 확인 | 비고 |
|---|---|---|---|---|---|---|---|---|---|
| 전문가 | 분야 | VARCHAR2 | 200 | | 15,976 | | 12 | √ | |
| 전문가 | 소속기관 | VARCHAR2 | 100 | | 15,976 | | 134 | √ | |
| 문헌 마스터 | 제목 | VARCHAR2 | 2000 | | 8,709 | | 1 | √ | |
| 물품 검수 | 검수자 코멘트 | VARCHAR2 | 255 | | 478,493 | 264 | 12 | √ | |
| 논문 | 저자명 | VARCHAR2 | 200 | | 33,362 | 34 | 1 | √ | |

또한 모든 컬럼이 NULL, 공백 또는 무의미한 값들로 채워 있어서 사용되지 않는 컬럼으로 보일 경우 별도로 표기한다.

〈표 3-23〉 미사용 컬럼 목록 예시

| 테이블명 | 컬럼명 | 데이터 타입 | 길이 | 기본 값 | 총건수 | 공백 수 | NULL 건수 | 미사용 컬럼확인 | 비고 |
|---|---|---|---|---|---|---|---|---|---|
| 첨부파일 | 파일확장자 | VARCHAR2 | 20 | | 189 | | 189 | √ | 미사용 컬럼 |
| 과제 마스터 | 삭제자 | VARCHAR2 | 40 | | 2,078 | | 2,078 | √ | 미사용 컬럼 |
| 과제 마스터 | 키워드 | VARCHAR2 | 1000 | | 287 | | 287 | √ | 미사용 컬럼 |
| 도서 마스터 | 플래그 | CHAR | 2 | | 287 | 287 | | √ | 미사용 컬럼 |

## 4.2 유일성 위반 내역

유일성 조건을 위반하는 것으로 추정되는 데이터의 개별 값 건수와 중복건수를 기재하여 목록을 작성한다.

〈표 3-24〉 유일성 위반 내역 목록 예시

| 테이블 명 | 컬럼 명 | 데이터타입 | 길이 | 총건수 | 개별값 건수 | 중복 건수 | 최대빈도값 | 개수 | 오류 확인 |
|---|---|---|---|---|---|---|---|---|---|
| 주문 마스터 | 주문코드 | VARCHAR | 12 | 408,231 | 402,103 | 6,128 | 00000 | 100 | √ |
| 고객 마스터 | 주민등록 번호 | VARCHAR | 13 | 301,123 | 300,001 | 1,122 | ############ | 13 | √ |
| 고객 마스터 | EMAIL | VARCHAR | 128 | 106,302 | 106,102 | 200 | abc@abc.com | 10 | √ |
| 고객 마스터 | 전화번호 | VARCHAR | 120 | 113,103 | 110,389 | 2,714 | 000-0000 | 100 | |

## 4.3 유효범위 위반 내역

유효 범위를 위반하는 것으로 추정되는 데이터의 유효범위 · 오류건수 · 총건수를 기재하여 목록을 작성한다.

〈표 3-25〉 유효범위 위반 내역 목록 예시

| 테이블 명 | 컬럼 명 | 데이터타입 | 길이 | 유효범위 | 총건수 | 오류건수 | 오류 확인 | 비고 |
|---|---|---|---|---|---|---|---|---|
| 부품 마스터 | 길이_XY | NUMBER | 22 | 0 이상 | 408,442 | 1 | √ | |
| 측정이력 | 방위각_도 | NUMBER | 22 | -360초과 360미만 | 333,154 | 10 | √ | |
| 측정이력 | 높이 | NUMBER | 22 | 0 이상 | 333,154 | 57 | √ | |
| 측정이력 | 고도 | NUMBER | 22 | 0 이상 | 333,154 | 47 | √ | |

## 4.4 허용값 위반 내역

허용값에 위배되는 것으로 추정되는 데이터의 개별값과 발생 빈도를 모두 나열하여 목록을 작성한다.

〈표 3-26〉 허용값 목록 위반 내역 예시

| 테이블명 | 컬럼명 | 데이터 타입 | 길이 | NULL 허용 | 총건수 | 데이터 | | | 비고 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 값 | 건수 | 유효 | |
| 수입품목 | 항목코드 | CHAR | 1 | Y | 10,047,409 | - | 1 | | |
| | | | | | | [NULL] | 557,547 | √ | |
| | | | | | | ] | 1 | | |
| | | | | | | - | 2 | | |
| | | | | | | 0 | 9,160,685 | √ | |
| | | | | | | 1 | 152,608 | √ | |
| | | | | | | 2 | 176,564 | √ | |
| | | | | | | 9 | 1 | √ | |
| 간행물 | 사용여부 | VARCHAR | 1 | N | 67,738 | [SPACE] | 533 | | |
| | | | | | | N | 66,283 | √ | |
| | | | | | | Y | 922 | √ | |
| 저장매체 | 매체코드 | CHAR | 1 | N | 10,047,409 | [NULL] | 557,547 | | |
| | | | | | | 0 | 9,439,289 | √ | |
| | | | | | | 1 | 30,221 | √ | |
| | | | | | | 2 | 20,343 | √ | |
| | | | | | | b | 9 | | |
| 보고서 | 언어명 | CHAR | 2 | N | 119,551 | KR | 109,457 | √ | |
| | | | | | | kor | 94 | | |
| | | | | | | EN | 10,000 | √ | |

## 4.5 문자열 패턴 내역

문자열 패턴 분석결과는 유효하지 못한 데이터 패턴으로 보이는 사례를 일부 발췌하여 패턴 목록과 건수를 기재하여 목록을 작성한다.

〈표 3-27〉 유효 문자열 패턴 목록 예시

| 테이블명 | 컬럼명 | 데이터 타입 | 길이 | NULL 허용 | 총건수 | 데이터 | | | 비고 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 유형 | 건수 | 유효 | |
| 교육장 | 관리 번호 | VARCHAR | 8 | Y | 13,317 | [NULL] | 5,213 | √ | |
| | | | | | | CCCCCC | 161 | | |
| | | | | | | CCCCC | 24 | | |
| | | | | | | 9 | 4,738 | √ | |
| | | | | | | 99 | 1 | √ | |
| | | | | | | 9-9 | 3,180 | √ | |
| 고객 마스터 | 우편번호 | CHAR | 7 | Y | 10,832 | [NULL] | 10 | √ | |
| | | | | | | 999SSS | 322 | | |
| | | | | | | 999SS | 4 | | |
| | | | | | | 999- | 1 | | |
| | | | | | | -999S | 1 | | |
| | | | | | | 999-999 | 10,494 | √ | |

## 4.6 날짜 유형 위반 내역

날짜유형의 분석결과는 유효날짜형식과 오류발생건수, 오류발생 날짜패턴 및 건수를 기재하여 목록을 작성한다.

〈표 3-28〉 날짜유형 분석 목록 예시

| 테이블명 | 컬럼명 | 데이터 타입 | 길이 | NULL 허용 | 총건수 | 오류 건수 | 유효날짜 형식 | 데이터 | | 비고 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 오류패턴 | 건수 | |
| 대출 신청 | 신청일 | CHAR | 8 | N | 13,317 | 167 | YYYYMMDD | 9999-99-99 | 129 | 9999-99-99형식은 응용프로그램에서 처리 |
| | | | | | | | | 999999 | 14 | |
| | | | | | | | | 9999 | 22 | |
| | | | | | | | | 99 | 2 | |
| | 지급일 | CHAR | 8 | Y | 13,317 | 130 | YYYYMMDD | 9999-99-99 | 112 | |
| | | | | | | | | 99/99/99 | 12 | |
| | | | | | | | | 9999 | 3 | |
| | | | | | | | | 9 | 3 | |

## 4.7 구조 무결성 위반 내역

구조분석 결과는 사전 정의된 관계 목록 · 기수성 · 진단건수 · 오류건수 등을 기재하여 목록으로 작성한다. 구조 분석은 일반 테이블 관계분석과 코드 분석을 구분하여 작성한다.

〈표 3-29〉 관계 분석 위반 내역 목록 예시

| 구분 | 기준 테이블명 | 기준 컬럼명 | 대상 테이블명 | 대상 컬럼명 | 대상 키컬럼명 | 기수성 | 진단 건수 | 오류 건수 | 오류율 (%) | 비고 |
|---|---|---|---|---|---|---|---|---|---|---|
| 관계 | MEMBER | MEMBER_NO | MEMBER_DTL | MEMBER_NO | MEMBER_NO | 1:M | 12,001 | 21 | 0.17 | |
| | PROJECT | PRJ_NO | PROJECT_SPLY | PRJ_NO | SPLY_ID | 1:M | 10,201 | 12 | 0.12 | |
| | PRODUCT | PRODUCT_CD | REPAIR_DTL | PRODUCT_CD | REPAIR_SN | 1:M | 1,221 | 212 | 17.36 | |
| | CUSTOMER | CUSTOMER_ID | ORDER | CUSTOMER_ID | ORDER_ID | 1:M | 23,21 | 21 | 0.90 | |

〈표 3-30〉 표준 코드 일관성 위반 내역 목록 예시

| 구분 | 코드 테이블명 | 코드 컬럼명 | 대상 테이블명 | 대상 컬럼명 | 키 컬럼명 | Null 허용 | 코드 조건 | 진단 건수 | 오류 건수 | 오류율 (%) | 비고 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 코드 | CODE_MASTER | CD | PROJECT | DEPT_CD | PROJECT_ID | Y | TYPE_CD='003' | 11,032 | 122 | 1.11 | |
| | CODE_MASTER | CD | PAPER | KIND_GB | PAPAER_ID | N | TYPE_CD='004' | 120,102 | 20 | 0.02 | |
| | CODE_MASTER | CD | SUPPLIER | WRK_GB | SUPPLIER_NO | Y | TYPE_CD='014' | 123,221 | 502 | 0.41 | |
| | CODE_MASTER | CD | SUPPORT | SPRT_GB | SUPPORT_NO | N | TYPE_CD='027' | 321,221 | 306 | 0.10 | |
| | CODE_MASTER | CD | ORDER | CMPN_GB | ORDER_NO | N | TYPE_CD='033' | 212,112 | 12 | 0.01 | |

## 4.8 프로파일링 결과서 작성

최종 확인된 프로파일링 결과서를 각 해당 업무 담당자들로부터 회수하여 오류데이터 발생 컬럼을 해당 업무별 · 테이블별 · 오류발생 유형별로 정리하여 최종 프로파일링 결과 종합 보고서를 작성한다. 또한 프로파일링 결과를 데이터 품질기준에 매핑하여 각 품질기준 유형별로 오류발생 추이를 파악할 수 있도록 종합 보고서를 작성한다.

〈표 3-31〉 프로파일링 결과 보고서 예시

프로파일링 결과 보고서

| 분석유형 | 테이블명 | 엔티티명 | 컬럼명 | 속성명 | 타입 | 길이 | Null | PK | 총건수 | Unique | 비고 | 값 | 건수 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 컬럼 | CSTMR_MST | 고객마스터 | KOI_NO | 관리번호 | VARSHAR2 | 30 | N | Y | 97,530 | 97,530 | Null 및 중복 없음. | - | - |
| 패턴 | CSTMR_MST | 고객마스터 | KOI_NO | 관리번호 | VARSHAR2 | 30 | N | Y | 97,530 | 6 | 6가지 패턴이 존재함. 모두 맞는 패턴인지 확인 필요. | SS99999999999 | 1 |
| | | | | | | | | | | | | S999999-99999 | 6 |
| | | | | | | | | | | | | S999999S99999 | 1 |
| | | | | | | | | | | | | S999999999999 | 97,353 |
| | | | | | | | | | | | | S999999`99999 | 1 |
| | | | | | | | | | | | | S999999S99999 | 168 |
| 컬럼 | CSTMR_MST | 고객마스터 | PR_NO | 주민번호 | VARSHAR2 | 14 | N | | 97,530 | 97,529 | 640605-1234567 2건 존재. 확인 필요. | 640605-1234557 | 2 |
| 패턴 | CSTMR_MST | 고객마스터 | PR_NO | 주민번호 | VARSHAR2 | 14 | N | | 97,530 | 12 | 6가지 패턴이 존재함. 96,040건은 유효한 패턴. 나머지 패턴을 갖는 데이터 확인 필요. | S99999-9999999 | 1 |
| | | | | | | | | | | | | 999999--999999 | 6 |
| | | | | | | | | | | | | 999999-S999999 | 1 |
| | | | | | | | | | | | | 999999-999 | 1 |
| | | | | | | | | | | | | 999999-99999 | 1 |
| | | | | | | | | | | | | 999999-999999 | 51 |
| | | | | | | | | | | | | 999999-9999999 | 96,040 |
| 컬럼 | CSTMR_MST | 고객마스터 | ID | 식별번호 | VARSHAR2 | 50 | Y | | 97,530 | 97,504 | Null이 26건 존재함. Null이 유효한지 확인 필요. | Null | 26 |
| 컬럼 | CSTMR_MST | 고객마스터 | SHI_NM | 한문성명 | VARSHAR2 | 32 | Y | | 97,530 | 47,525 | 45,562건(46.72%)이 Null임. 컬럼 특성상 Null일 가능성 높음. | Null | 45,562 |
| 패턴 | CSTMR_MST | 고객마스터 | SHI_NM | 한문성명 | VARSHAR2 | 32 | Y | | 97,530 | 65 | 65개의 필요 이상의 다양한 패턴이 존재함. 확인 필요. | 𘚟𘚟𘚟 | 1 |
| | | | | | | | | | | | | HBBBBBBBBBBBBBBBBBBBBBBBBH | 1 |
| | | | | | | | | | | | | SSS,SSSS-SSSS | 1 |
| | | | | | | | | | | | | HBHB𘚟 | 1 |
| | | | | | | | | | | | | HHH | 42,599 |
| | | | | | | | | | | | | Null | 45,560 |
| | | | | | | | | | | | | 이외 65개 패턴 존재 | |
| 컬럼 | CSTMR_MST | 고객마스터 | ENG_NM | 영문이름 | VARSHAR2 | 64 | Y | | 97,530 | 59,493 | 30,499건(31.27%)의 Null 존재. 확인 필요. | Null | 30,499 |
| 컬럼 | CSTMR_MST | 고객마스터 | HOME_POST_NO | 우편번호 | VARSHAR2 | 7 | Y | | 97,530 | 13,475 | 42,867건(43.95%)의 Null 존재. 확인 필요. | Null | 42,867 |
| 패턴 | CSTMR_MST | 고객마스터 | HOME_POST_NO | 우편번호 | VARSHAR2 | 7 | Y | | 97,530 | 8 | 8개의 패턴이 존재함. 999-999외의 패턴은 오류로 추정. 확인 필요. | - | 1 |
| | | | | | | | | | | | | -- | 2 |
| | | | | | | | | | | | | 999 | 1 |
| | | | | | | | | | | | | 999- | 1 |
| | | | | | | | | | | | | 999-99 | 9 |
| | | | | | | | | | | | | 999-999 | 45,541 |
| | | | | | | | | | | | | 99999 | 62 |
| | | | | | | | | | | | | 999999 | 9,046 |

# 제2절 업무규칙

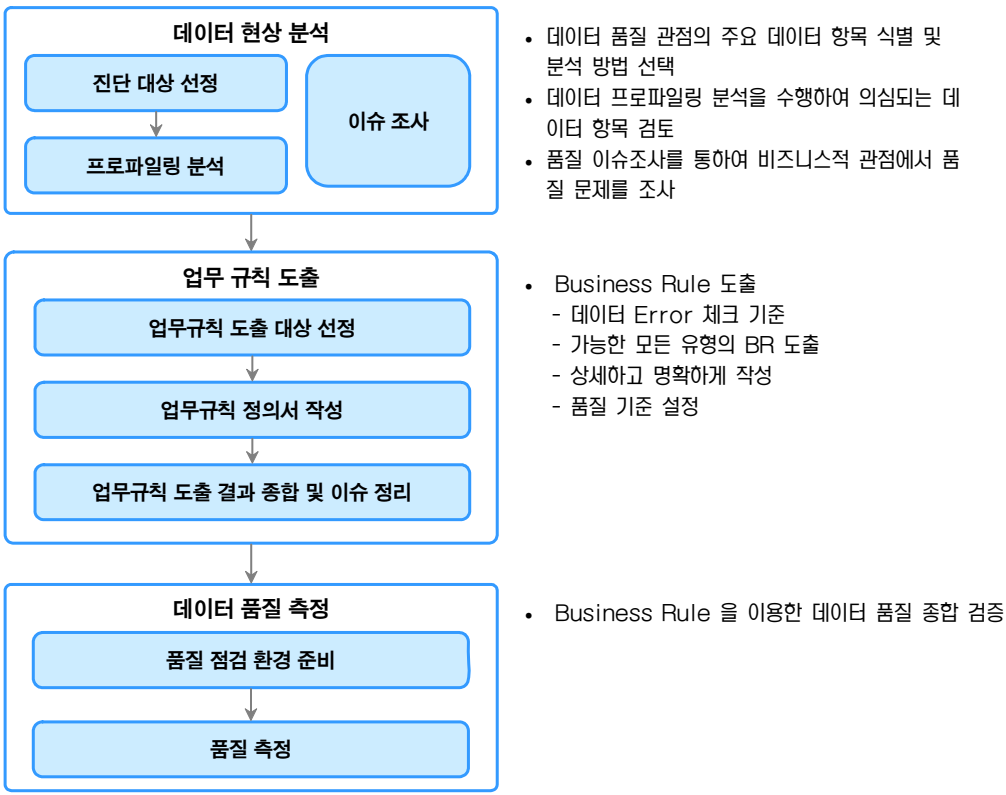
업무규칙은 업무와 관련된 모든 데이터의 규칙을 말한다. 업무규칙은 조직에서 데이터 품질을 지속적으로 관리하기 위해 사용하는 데이터 측정 규칙이며 데이터 값이 정확하기 위한 조건 표현이다.

업무규칙은 데이터 관리문서에 명시된 데이터 규칙, 업무담당자가 지식으로 알고 있는 규칙, 응용프로그램에 코딩된 규칙 등의 산재되어 분산 관리되고 있는 규칙을 통합한 것으로 조직 내부에서 운영하고 있는 정보시스템의 품질을 지속적으로 관리하기 위해 활용된다.

업무규칙은 업무 데이터와 관련된 모든 조직에 내재되어 있으며, 해당 조직원이 그 중요성을 인식해야 한다. 따라서 업무규칙은 업무 전문가와 품질 분석가들이 업무분석과 관리문서의 검토 등의 공동 작업을 통하여 별도로 도출한다.

## 1. 업무규칙 도출 절차

데이터 품질 이슈 조사 및 데이터 현상 분석 단계를 통하여 핵심 업무 테이블 및 컬럼을 토대로 업무규칙을 도출하여 품질측정 단계에서 활용한다. 다음은 업무규칙 도출 절차에 대한 일반 절차를 나타낸다.




〈그림 3-12〉 업무규칙 도출 절차

데이터 현상을 분석하는 방법은 정보시스템 내부이용자와 외부 고객의 이슈로부터 데이터 품질을 파악하는 접근방식과 데이터소스를 분석하여 오류현상 및 문제점을 발견해 나가는 방식으로 구분된다. 전자는 설문 및 담당자 인터뷰 등을 통한 품질 이슈조사를 통하여 이루어지며, 후자는 데이터 프로파일링 분석을 통하여 이루어진다.

사전에 파악된 데이터 품질 이슈와 프로파일링 결과물을 토대로 업무규칙을 도출하여 이를 지속적 품질 유지를 위한 주요 관리 대상으로 선정한다. 업무규칙 도출 작업의 효율성을 높이기 위하여 업무규칙 도출 대상 항목을 우선 선정한 후

해당 데이터 항목을 토대로 업무규칙 정의서를 작성한다. 작성된 업무규칙 정의서는 취합되어 최종 검토 작업을 수행한 이후 최종 공식 문서화 한다.

| 프로파일링 결과 | 오류추정 데이터 분석 | 업무규칙 도출 |
|---|---|---|
| 1. 인력 기본정보 테이블의 주민번호 중 잘못된 주민번호(999999-9999999, 000000-000000)가 존재 | 과거 데이터 중에서는 구분 컬럼이 외국인일 경우 임의 입력된 데이터로 '외국인 여부'가 Y일 경우는 주민번호가 잘못될 수 있음 | 인력기본정보 테이블의 주민번호는 '외국인 여부'가 N일 경우 유효함 |
| 2. 추가적으로 도출해야할 업무규칙은? | 주민번호의 번호체계로 볼 때, 앞번호는 생년월일과 일치해야 하며, 뒷번호 맨 첫자리는 남녀 성별과 일치해야 함<br><br>주민번호 뒷번호의 맨 첫자리는 1900~1999년 출생이고, 성별이 남성일 경우 '1'이어야 하며, 여성일 경우 '2'를 만족해야 한다. 또한 2000년 이후 출생일 경우는 남성일 경우 '3', 여성일 경우는 '4'를 만족해야 함<br><br>인력기본정보의 주민번호는 동일한 인력관리번호일 때 연구실적 테이블의 주민번호와 동일해야 함 | 주민번호 앞 여섯자리는 생년월일과 일치해야 함<br><br>주민번호 뒷번호 맨 첫자리는 출생년도 및 성별에 따른 생성규칙을 준수해야 한다<br><br>인력기본정보와 연구실적 두 테이블의 인력관리번호가 같으면, 주민번호도 동일해야 함 |

〈그림 3-13〉 프로파일링 결과로부터 추가 업무규칙 도출 과정 사례

작성된 업무규칙은 조직에서 지속적 품질 개선을 위하여 변경 및 개정 관리를 수행하며 업무규칙의 측정 위치, 용도를 결정하여 측정 주기에 따라 상시 데이터 품질을 측정하고 모니터링을 수행한다.

### 1.1 업무규칙 도출 대상 선정

업무규칙을 도출할 대상 데이터 항목을 사전에 정의하고, 작업의 순서를 정하여 계층적으로 수행함으로써 업무규칙 도출 시 누락된 항목을 최소화하고 작업의 효율을 높일 수 있다.

업무규칙 도출 대상을 다음과 같이 선정한다.

1) 데이터 관리문서 · 테이블 및 컬럼 중요도 · 프로파일링 결과서 · 품질이슈 정의서 · 핵심 데이터 항목 정의서 등 사전에 수행된 작업을 토대로 자료를 수집하여 업무규칙 도출 대상 목록을 작성한다.

〈표 3-32〉는 업무규칙을 도출해야할 컬럼 목록의 예이다. 작성된 컬럼 목록과 프로파일링 결과를 토대로 업무규칙으로 관리해야 할 컬럼 후보를 결정한다. 또한 일반적으로 제시된 품질기준 이외의 복수 컬럼간의 값의 선순위 파악이나 값이 정확하기 위한 조건을 비고란에 기재하여 복잡한 데이터 규칙을 도출하는데 활용한다.

〈표 3-32〉 업무규칙 도출 대상 컬럼 목록(단일 컬럼)

| 영문 테이블명 | 한글 테이블명 | 영문 컬럼명 | 한글 컬럼명 | 프로파일링 결과(오류건수) | | | | | | 핵심 정보 항목 | 업무규칙 관리 대상 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 총건수 | 완전성 | 유효성 | | | 유일성 | 소계 | | 완전성 | 유효성 | | | 유일성 | 비고 |
| | | | | | | 날짜 | 범위 | 포맷 | | | | | 날짜 | 범위 | 포맷 | | |
| PROJECT | 사업 | PRJ_NO | 사업번호 | 13,211 | 11 | | | 103 | 14 | 114 | | √ | | | √ | √ | |
| | | PRJ_NM | 사업명 | 13,211 | 102 | | | | | | √ | √ | | | | | |
| | | DEPT_CD | 사업부처 | 13,211 | | | | 25 | | 25 | | | | | √ | | |
| | | START_DT | 사업시작일 | 13,211 | | 210 | | | | 210 | √ | | √ | | | | 시작일은 종료일이전 |
| | | END_DT | 사업종료일 | 13,211 | | 132 | | | | 132 | √ | | √ | | | | |
| PAPER | 논문 | KIND_GB | 학술지구분 | 25,215 | 132 | | | 22 | | 154 | √ | √ | | | √ | | 구분이 '03' 이면 .. |
| | | ANNNC_DT | 논문발표일 | 25,215 | | 144 | | | | 144 | | | √ | | | | |
| | | VOL_NM | 게재권/집 | 25,215 | | | | | | | | | | | | | |
| | | NUM_NO | 게재호 | 25,215 | | | 121 | | | 121 | | | | √ | | | |
| | | START_PG | 시작페이지 | 25,215 | | | 131 | | | 131 | | | | √ | | | |
| | | END_PG | 종료페이지 | 25,215 | | | 23 | | | 23 | | | | √ | | | |

2) 작성된 대상목록을 토대로 업무규칙 도출대상을 선정한다. 데이터 프로파일링 결과 오류가 발생하는 대상을 위주로 해당 테이블 및 컬럼의 중요도 · 품질이슈 대상항목 · 핵심 정보항목 등을 고려하여 업무규칙 도출 순서를 정하여 업무규칙 도출 대상을 선정한다.

- 데이터 활용도 및 중요도에 따라서 업무규칙을 도출할 테이블의 작업 순서를 결정한다. 업무별 작업순서는 공통 업무를 기준으로 하여 관련 업무의 중요도와 의존도에 따라서 확장한다.

  예) 공통업무 → 고객관리 → 거래관리....

- 테이블의 작업 순서는 전사 표준코드 테이블 · 마스터 · 주요결과 · 집계 · 이력 · 단순 참조의 수순으로 계층적으로 수행한다.
- 동일한 업무 항목간 표준코드 테이블 · 표준도메인 · 도메인별 유효범위 등을 사전에 정의하여 동일한 항목의 유효성 판단 시 활용하도록 한다.
- 업무규칙 도출 대상 항목의 순서는 단일 항목에서 복수 항목 및 복수 테이블로 확장한다.

  예) 필수 값의 누락 → 유효범위 및 허용값 → 테이블간의 관계 → 복잡한 업무관계

3) 업무규칙을 적용할 데이터 항목별 측정 기준 · 용도 · 주기 · 핵심 데이터 여부 · 해당 중요도 등을 결정한다.

- 각 항목별 측정해야할 품질기준을 선정한다.
- 해당 데이터 항목에도 도출 가능한 모든 유형의 업무규칙을 도출해야 한다.
- 업무규칙을 적용할 용도를 정의한다.

  - 트랜잭션 처리 용도(데이터의 입력 · 수정 · 삭제 등)

- 데이터 이동 및 재구조화(추출 · 정제 · 변환 · 탑재 · 통합 등)

• 적용할 업무규칙의 측정주기 및 시기를 고려하며 업무절차별 트랜잭션 발생시  수행할 것인지 혹은 각 측정 주기에 따라 일별 · 주별 · 월별로 수행할 것인지를 정의한다.

4) 도출된 대상 데이터 항목 중에서 누락되었거나 혹은 중복된 데이터 항목이 있는지 점검한다. 단일 데이터 항목이 자료사전과 공통코드 테이블에 정의되었는지 재차 점검하고 누락되거나 중복된 사항이 발견되면 정정한다.

## 1.2 업무규칙 정의서 작성

사전에 정의된 업무규칙 도출 대상 항목으로부터 해당 품질기준과 데이터 항목과 매핑되는 업무규칙을 도출하여 업무규칙 정의서를 작성한다. 업무규칙은 데이터의 품질관리를 위해 지속적으로 관리되어야 하는 데이터의 규칙으로 데이터의 값이 정확하기 위한 조건에 대한 표현이다. 따라서 업무규칙은 일관되고 정형화된 양식으로 작성된다.

• 업무규칙은 관련 업무에 대한 구분이나 업무명 등을 기재한다.

• 업무규칙은 테이블의 컬럼 단위로 작성한다.

• 업무규칙은 자연어의 형태로 자유롭게 기술하며 한글테이블 및 컬럼, 영문테이블 및 영문 컬럼을 모두 명시하여 업무담당자 및 IT 담당자가 모두 이해하기 쉽게 작성한다.

• 다수의 테이블과 컬럼과 관련이 있는 업무규칙의 경우 데이터의 측정 대상과 관련된 컬럼을 선정한다.

- 해당 컬럼이 별도의 코드나 ID를 사용하는 경우 관련 코드의 유효 값과 의미를 모두 기재한다.
- 업무규칙은 반드시 해당 품질기준을 명시한다. 품질기준은 품질측정 전에 사전에 정의한 품질기준을 활용한다.

〈표 3-33〉 업무규칙 정의서 작성 사례

<table>
<tr><th colspan="8">업무규칙 정의서</th></tr>
<tr><td>업무규칙ID</td><td colspan="7">BR20080103</td></tr>
<tr><td>업무구분</td><td colspan="2">기금관리</td><td>업무명</td><td colspan="3">연구기금지급</td></tr>
<tr><td>업무규칙</td><td colspan="7">기금유형 코드(TB_FUND.FUND_TYPE)가 '08' 이면 지급일(SUPPLY_DT)은 NULL 이다<br>(08 : 미지급처리)</td></tr>
<tr><td>데이터베이스명</td><td colspan="7">기금관리 DB</td></tr>
<tr><td>대상테이블</td><td colspan="2">지급내역(TB_FUND)</td><td>대상 컬럼</td><td colspan="3">지급일(TB_FUND.SUPPLY_DT,)</td></tr>
<tr><td>관련 테이블명</td><td colspan="2"></td><td>관련 컬럼명</td><td colspan="3">기금유형(TB_FUND.FUND_TYPE)<br>지급내역코드(TB_FUND.FUND_CD)</td></tr>
<tr><td>품질기준</td><td colspan="2">정확성</td><td>세부품질기준</td><td colspan="3">업무규칙 정확성</td></tr>
<tr><td rowspan="2">중요도</td><td>테이블중요도</td><td>컬럼<br>중요도</td><td>시스템<br>연관성</td><td>서비스<br>영향도</td><td>업무<br>영향도</td><td colspan="2">중요도 계</td></tr>
<tr><td>A</td><td>A</td><td>A</td><td>B</td><td>B</td><td colspan="2">3.5</td></tr>
<tr><td colspan="8">측정 스크립트</td></tr>
<tr><td>총건수 SQL</td><td colspan="7">SELECT COUNT(*) AS TOT_CNT<br>FROM TB_FUND</td></tr>
<tr><td>오류건수 SQL</td><td colspan="7">SELECT COUNT(*) AS ERR_CNT<br>WHERE FUND_CD = '08' AND<br>TB_FUND.SUPPLY_DT IS NOT NULL</td></tr>
<tr><td>오류추출 SQL</td><td colspan="7">SELECT TB_FUND.FUND_TYPE, TB_FUND.SYPPLY_DT, TB_FUND_CD<br>FROM TB_FUND<br>WHERE FUND_CD = '08' AND<br>TB_FUND.SUPPLY_DT IS NOT NULL</td></tr>
<tr><td>작성자</td><td colspan="2">김관리</td><td>작성일자</td><td>2008.01.01</td><td>버전</td><td>1.2</td></tr>
</table>

- 업무규칙의 중요도는 업무규칙이 해당 업무와 서비스에 미치는 영향, 품질기준의 중요도 등을 사전에 선정하여 각각 비중에 맞게 배분하여 계산한 후 기입한다.
- 측정 SQL은 현재 운영 중인 DBMS의 종류와 버전에 맞는 문법으로 작성을 하며 오류데이터를 추출하는 부정적 문법으로 작성한다. 해당 테이블의 총건수와 오류건수를 출력하는 SQL을 같이 관리하면 향후 오류율 추출 시 효과적이다.
- 향후 관리목적으로 작성자 · 작성일자 · 업무규칙의 관리버전 등을 기재한다.

## 1.3 업무규칙 조정 및 확정

작성된 업무규칙은 다수의 업무전문가 · 담당자 · 품질 분석가 등이 공동 작업으로 도출하는 특성상 도출된 업무규칙의 용어, 표현이 일관되지 못하거나 작업자간의 업무 데이터에 관한 이해 또는 견해 차이로 적절하지 못하거나 중복된 업무규칙이 도출될 수 있다. 따라서 도출된 업무규칙을 확인하고 미비사항이 발견될 경우 보완해야 할 필요가 있다. 업무규칙 보완 작업 시 확인할 사항은 다음과 같다.

- 중요 업무규칙의 누락이 없는지 파악한다.
- 업무규칙의 중복성이 있는지 파악한다.
- 업무규칙이 정확히 도출되었는지 파악한다.
- 동일 항목이 업무 별로 다르게 작성된 업무규칙이 있는지 파악한다.
- 유사한 성격의 항목이 전혀 다르게 업무규칙으로 작성된 경우가 있는지 파악한다.

- 업무규칙 표현이 업무규칙 작성 표준과 일치하고 일관된 표현을 사용하는지 파악한다.
- 가독성이 떨어지거나 난해한 표현으로 기술되었는지 파악한다.

보완된 업무규칙 정의서를 모두 취합하여 확정한다. 확정된 업무규칙은 결재권자의 승인을 얻어 공식 문서화 한다. 업무규칙은 데이터의 개정, 변경 등의 사유로 해당 규칙의 변경이 요구될 때 그 개정 및 변경 이력을 모두 관리한다.

## 2. 업무규칙 작성 지침

업무규칙은 데이터 품질측정을 위한 오류율 측정 및 오류 사례 도출을 주된 목적으로 한다. 따라서 규칙을 적용할 컬럼의 오류데이터를 찾기 위한 데이터 규칙을 도출하며 Not True의 관점에서 오류데이터 추출을 위한 조건을 작성한다. 업무규칙은 업무규칙 정의와 오류데이터 검증 규칙의 두 가지 유형으로 구분된다. 업무규칙 정의는 자연어로 값이 진실이어야 하는 데이터 규칙을 조건으로 기술하며 오류데이터 검증 규칙은 오류데이터를 추출하기 위한 논리적 언어로 표현된 SQL문 또는 조건식으로 구성된다. 업무규칙 작성 시 고려할 사항은 다음과 같다.

- 업무규칙의 도출은 컬럼 단위로 작성한다.
- 업무규칙은 데이터 품질측정을 위한 오류데이터 검증을 위한 규칙을 도출한다.
- 진단 대상 데이터 항목과 관련된 업무규칙을 도출할 때 도출 가능한 모든 유형의 업무규칙을 각각 도출한다.
- 다수의 데이터 항목과 관련 있는 업무규칙은 측정 결과를 나타내는 데이터 항목을 업무규칙의 측정 대상으로 지정

한다. 결과를 나타내는 데이터 항목을 구분하기 어려운 경우에는 업무적으로 더 중요한 항목을 선정한다.

- 공통 코드에 등록된 코드와 범위 값은 필히 공통 코드 데이터를 참조하도록 한다. 공통 코드에 등록이 되지 않은 코드와 범위 값은 이슈 대상 리스트로 취합하여 향후 공통 코드에 추가 등록해야 된다.

〈표 3-34〉 업무규칙별 진단 대상 항목 선정 예시

| |
|---|
| 업무규칙#1 : 정보조회의뢰내역.상태(TB001.CRNT_STTS)는 '001'일 경우 의뢰자ID(TB001.USER_ID)는 반드시 존재하여야 한다.<br>진단 대상 항목 : **의뢰자 ID** |
| 업무규칙#2 : 제품분할여부(TB_SPT_F)가 'Y' 이고 제품분할수(TB_SPT_CNT)가 '0' 보다 크면 제품분할수(TB_SPT_CNT)에 존재하는 수량만큼 분할하기 이전의 재료번호가 동일하게 존재해야 한다.<br>진단 대상 항목 : **재료번호** |

- 공통 코드에 등록하지 않고 별도로 임의 코드를 부여하여 사용하는 ID 성격의 컬럼은 해당 유효 값을 모두 나열한다.

〈표 3-35〉 개별코드 사용 업무규칙 작성 예시

| |
|---|
| 업무규칙#3 : 환불내역.처리구분(TB0003.PRCS_GB)은 0:환불대기, 1:환불처리중, 2:환불처리완료, 3:미처리대상 값을 가져야 한다. |

- 업무규칙 정의 시 데이터 항목명은 필히 한글명과 영문명를 이용하여 상세하게 작성하며 진단 대상 테이블 및 컬

럼명, 관련 테이블 및 컬럼명을 모두 명시하여 가독성을 높이도록 한다.

- 업무규칙을 작성 시 개별 데이터 항목에 적용한 데이터 품질기준이 서로 독립적이어야 하며 서로의 중복 측정이 배제 되어야 한다. 예컨대 유효성 · 유일성 · 일관성 등의 품질기준에 해당되는 데이터 항목의 업무규칙 SQL을 작성할 때에는 값의 완전성 조건(NOT NULL)을 항상 명시해야 된다. 이미 NULL 값이 발생하여 오류로 확정된 데이터 항목을 타 품질기준을 적용하여 측정할 때에 NULL 값이 오류로 측정이 되어 오류율의 중복이 발생될 수 있다. 따라서 업무규칙을 작성할 때는 값의 필수여부, 기본값 등을 사전에 파악하여 명확히 작성해야 한다. 이러한 사항은 다수의 업무테이블 간의 관계를 분석하거나 상관성분석을 수행할 때에도 공통으로 고려한다.

## 3. 업무규칙 및 BR-SQL 사례

복수컬럼의 누락관계 · 날짜의 선순위 관계 · 유도된 컬럼의 계산관계 · 복수컬럼의 조합으로 발생하는 데이터 중복 등을 기준으로 업무규칙을 작성하는 BR SQL을 소개한다. BR (Business Rule) SQL은 업무규칙을 검증하기 위한 스크립트로서 업무규칙에 위배되는 데이터를 추출하기 위한 SQL문을 의미한다. 따라서 BR-SQL은 오류데이터를 추출하기 위한 부정형으로 작성된다. 도출된 업무규칙은 SQL문으로 변경 시 작성된 업무규칙이 지켜져야 하는 긍정형의 규칙을 NOT TRUE의 형태로 전환하여 작성한다.

## 3.1 날짜 및 시간의 선순위 관계

| PRJ_NO | YEAR | ORG_CD | DEPT_CD | PRJ_NM | START_DT | END_DT | |
|---|---|---|---|---|---|---|---|
| 200608101 | 2006 | 2034 | 03 | 멀티미디어 통신 전송 품질 제어 | 2006/01/01 | 2006/12/31 | |
| 200608102 | 2006 | 3042 | 03 | 네트워크망을 이용한 품질 시스템 | 2006/03/01 | 2005/10/20 | ① |
| 200703103 | 2007 | 1043 | 04 | 품질관리 서버 구축 | 2006/01/01 | 2007/03/01 | |
| 200704123 | 2007 | 1043 | 04 | 품질관리 클라이언트 시스템 구축 | 2007/03/05 | 2007/12/31 | |
| 200704124 | 2007 | 1001 | 04 | 차세대 통합 품질 관리 시스템 | | 2007/09/20 | ② |
| 200704125 | 2007 | 1001 | 05 | 무선 통신 통화품질 향상 | 2007/04/21 | 2006/12/31 | |
| 200704126 | 2007 | 1030 | 04 | 6시그마 품질 경영 시스템 | 2008/03/01 | 2007/10/03 | ③ |
| 200704127 | 2007 | 1030 | 05 | 생산라인 품질관리 시스템 | 2007/04/01 | | |

〈그림 3-14〉 날짜 선순위 분석 대상 목록

〈그림 3-14〉에서 날짜를 표현하는 값은 업무의 시작·종료와 같이 그 컬럼이 내포하는 의미가 단순히 시작일과 종료일의 값을 저장하여 활용하는 것만을 의미하지 않고, 날짜 컬럼에 값이 있고 없음에 따라서 시작했음과 종료했음의 업무 상태를 나타내는 데까지 활용된다.

따라서 ①의 경우에는 품질연구 프로젝트의 시작일과 종료일의 선순위 관계가 역행되어 있는 예이며, ②의 경우에는 시작일이 누락되었는데 종료일이 존재하는 경우에 해당된다. ①의 경우에는 명확히 날짜의 선순위 관계가 역행되어 있지만, ②와 같이 종료일은 있으되 시작일의 값이 누락된 경우는 시작하지도 않은 프로젝트가 이미 종료되었음을 나타내는 사례로 볼 수 있다. 값이 누락된 컬럼은 해당 컬럼이 업무에서 활용되는 목적과 기능에 따라 다양하게 해석할 수 있다. 날짜 및 시간의 선순위와 관련된 업무규칙의 위반사항을 검증하는 BR SQL은 다음과 같다.

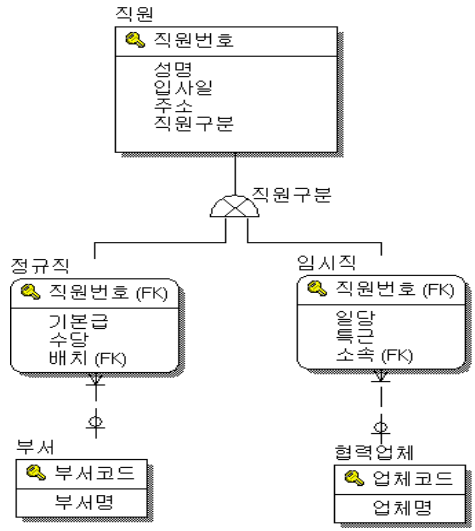
- 업무규칙#1 : 프로젝트의 시작일은 종료일보다 이전 시점이어야 한다.
- 업무규칙#2 : 프로젝트의 종료일이 있으면 시작일은 반드시 존재해야 한다.

〈표 3-36〉 날짜 선순위 관계 업무규칙 검증 SQL

```
/* 시작일이 종료일보다 이후 시점인 경우 */

SELECT START_DT, END_DT, PRJ_NO
FROM   PROJECT
WHERE (START_DT > END_DT) ;

/* 종료일이 있는데 시작일이 없는 경우 */

SELECT START_DT, END_DT, PRJ_NO
FROM   PROJECT
WHERE (START_DT IS NULL AND END_DT IS NOT NULL) ;
```

## 3.2 물리 모델링 시 서브타입 설계 반영

〈그림 3-15〉은 논리 모델에서의 직원 엔티티와 부서 및 협력 업체의 관계를 ERD로 표현한 예이다. 직원 엔티티를 살펴보면 업체는 직원의 구분에 따라 정규직 직원과 임시직 직원으로 구분된다. 또한 정규직 직원의 경우는 소속 부서가 배치되고 임시직 직원의 경우에는 소속 협력 업체가 지정된다.




〈그림 3-15〉 직원 엔티티 관련 ERD

이러한 업무 관계를 물리 데이터베이스에 설계 반영할 경우에는 업무 실체가 가지는 특성과 활용 목적에 따라 다양한 방법으로 구성하게 된다. 보통 하나의 테이블에 모든 속성을 포함시키거나 여러 개의 테이블로 분할하여 관계를 설정하도록 설계한다. 일반적으로 물리 모델을 데이터베이스에 반영할 때 해당 컬럼간 제약조건을 시스템에 설정해야 하나 실무에서는 확인 제약 조건을 생략하고 응용프로그램에서 검증 로직을 사용하는 경우가 많다.

| EMP_NO | NAME | EMP_TYPE | SAL | DEPT_CD | CO_ORG_CD | DAY_FEE | EXT_FEE |
|---|---|---|---|---|---|---|---|
| 2032 | 임수경 | R | 910000 | 11 | | 1000 | |
| 3541 | 강수전 | T | | 12 | 1000 | 35000 | 12000 |
| 5846 | 심민하 | T | | 15 | 1342 | 40000 | 10000 |
| 3541 | 김현준 | R | 990000 | 10 | | | |
| 1223 | 한준명 | R | 780000 | 14 | 1130 | | 30000 |

〈그림 3-16〉 서브타입을 하나의 테이블로 설계한 직원 테이블 예시

〈그림 3-16〉은 데이터베이스에 실제 설계가 반영된 직원 테이블이며 서브타입을 하나의 테이블로 설계한 예이다. 직원구분(EMP_TYPE)은 정규직(R)과 임시직(T)을 나타내며, 정규직인 직원은 협력업체코드(CO_ORG_CD)와 일당(DAY_FEE), 특근(EXT_FEE) 등의 값이 없어야 하나 '1130', '30000' 등의 오류데이터가 있다.

- 업무규칙#1 : 직원구분이 'R'(정규직)일 경우 일당은 NULL 이어야 한다.
- 업무규칙#2 : 직원구분이 'R'(정규직)일 경우 협력업체코드는 NULL 이어야 한다.

〈표 3-37〉 직원 급여 관련 업무규칙 SQL 예시

```
/* 정규직원인데 일당이 있는 경우 */
SELECT EMP_TYPE, DAY_FEE, EMP_NO
FROM   EMPLOYEE
WHERE EMP_TYPE = 'R' AND DAY_FEE IS NOT NULL ;

/* 정규직원인데 협력업체 코드가 있는 경우 */
SELECT EMP_TYPE, CO_ORG_CD, EMP_NO
FROM   EMPLOYEE
WHERE EMP_TYPE = 'R' AND CO_ORG_CD NOT NULL ;

/* 임시직원인데 기본급이 있는 경우 */
SELECT EMP_TYPE, SAL, EMP_NO
FROM   EMPLOYEE
WHERE EMP_TYPE = 'T' AND SAL IS NOT NULL ;
```

- 업무규칙#3 : 직원구분이 'R'(정규직)일 경우 특근은 NULL 이어야 한다.
- 업무규칙#4 : 직원구분이 'T'(임시직)일 경우 기본급은 NULL 이어야 한다.

## 3.3 유도된 컬럼 속성 값

유도 속성은 업무의 필요성에 의하여 도출된 속성으로서 하나 이상의 컬럼으로부터 값이 파생되거나 새롭게 도출이 된 속성 값을 말한다. 예컨대 고객의 연령, 학생의 재학기간, 고객의 신용도 등이 이에 해당된다. 유도 속성은 컬럼 값의 변경에 따라 타 속성에 대한 의존성이 높으므로 데이터베이스 구조 설계 시에는 직접 반영되지 않고 다수 컬럼으로 재계산하여 응용프로그램에서 활용하도록 설계하는 것이 바람직하나 성능 향상 또는 기타 관리 목적으로 설계하는 경우가 있다.
유도 속성 중에서는 몇 가지 컬럼과 계산식으로 쉽게 계산이 가능한 연령 · 재학기간 같은 속성도 있지만 보다 복잡한 고객의 신용도 · 재무 위험도 등의 다양한 집계 값과 별도의 조건식이 필요할 수 있다.

연령 = 현재년도 - 출생년도
재학기간 = 현재일 - 입학일

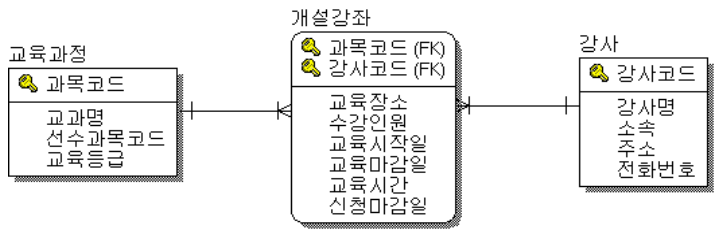
- 업무규칙 : 연령은 현재년도 - 출생년도를 계산한 값과 동일해야 한다.

〈표 3-38〉 유도 속성 관련 업무규칙 SQL 예시

```
SELECT BIRTH, AGE, STUDENT_NO
FROM   STUDENT
WHERE TO_CHAR(SYSDATE, 'YYYY') - TO_CHAR(BIRTH,'YYYY') <> AGE ;
```

## 3.4 복수컬럼의 유일성

복수 컬럼이 동시에 유일해야 하는 경우는 여러 가지 자연키의 조합으로 그 유일성을 입증할 수 있을 경우에 해당된다.




〈그림 3-17〉 교육과정 개설 ERD 예시

| COURSE_CD | LECTURER_CD | PLACE | APP_MNT | START_DT | END_DT |
|---|---|---|---|---|---|
| C071012 | L209 | 1강의실 | 10 | 2007-10-15 | 2007-10-19 |
| C071012 | L207 | 2강의실 | 12 | 2007-10-16 | 2007-10-20 |
| C071012 | L230 | 3강의실 | 30 | 2007-10-17 | 2007-10-21 |
| C071013 | T102 | 3강의실 | 54 | 2007-10-18 | 2007-10-22 |
| C071013 | T102 | 3강의실 | 34 | 2007-10-18 | 2007-10-22 |
| C071015 | T104 | 2강의실 | 36 | 2007-10-20 | 2007-10-24 |
| C071015 | L020 | 3강의실 | 27 | 2007-10-21 | 2007-10-25 |
| C071015 | L021 | 6강의실 | 35 | 2007-10-22 | 2007-10-26 |

〈그림 3-18〉 개설 강좌 테이블 예시

〈그림 3-18〉은 교육과정과 강사와의 관계로 구성이 되는 개설 강좌 테이블이다. 개설강좌는 과목코드와 강사코드, 교육장소가 모두 동일한 데이터가 발생되어서는 안 된다. 이 사례는 복수 컬럼의 조합으로 이루어진 컬럼의 유일성 검증 절차이며 BR-SQL은 다음과 같다.

- 업무규칙 : 교육과정코드 · 강사코드 · 교육장소가 모두 조합되었을 경우 유일한 값을 가져야 한다.

〈표 3-39〉 복수컬럼 유일성 관련 업무규칙 SQL 예시

```
SELECT COURSE_CD, LECTURER_CD, PLACE
FROM   STUDENT
GROUP BY COURSE_CD, LECTURER_CD, PLACE
HAVING COUNT(*) > 1 ;
```

복수 컬럼의 조합과 관련된 유일성은 특정 시점에서 공유될 수 없는 상황과 연관되며 동일 장소와 시간에 특정 주체가 동시에 발생할 수 없는 특징을 가진다. 위의 대표적인 사례로는 도서의 대출, 회의실의 예약, 자동차의 임대 등이 해당된다. 특정 개체들이 동시에 같은 시간대에 중복된 데이터가 존재할 경우 업무상 중복이 되거나 혼선을 일으킬 수 있다. 이러한 사례도 위의 SQL을 변경하여 오류데이터의 검증이 가능하다.

## 3.5 계산 및 집계 관계

다수의 테이블의 값을 집계함수를 활용하여 실시간으로 통계정보를 서비스하는 시스템을 고려해볼 때 서비스 요청시마다 필요한 데이터를 추출하여 서비스하는 것은 매우 비효율적이

다. 따라서 특정 통계 테이블을 두거나 중복 테이블을 추가하여 활용할 수 있다.

〈그림 3-19〉는 주문 완료된 데이터의 합을 계산하여 월간 매출 통계를 입력하는 테이블을 나타낸 예이다. 월간통계(MONTHLY_STATISTICS) 테이블의 해당 년월의 합과 주문 상세(ORDER_DETAIL)테이블의 합을 서로 비교하여 불일치하는 값(2008년 2월의 총수입)을 추출할 수 있다.

**MONTYLY_STATISTICS**

| YEAR | MONTH | ORDER_CNT | CANCEL_CNT | TOTAL_INCOME |
|---|---|---|---|---|
| 2007 | 12 | 2 | 0 | 66000 |
| 2008 | 01 | 3 | 1 | 50000 |
| 2008 | 02 | 2 | 2 | 60000 |

**ORDER_DETAIL**

| ORDER_NO | ORDER_DT | ORDER_TOT_AMT | STATUS | APPRV_TYPE | SEND_DT | COMPLETE_DT |
|---|---|---|---|---|---|---|
| 1 | 2007-12-05 | 30000 | 배송완료 | 카드 | 2007-12-10 | 2007-12-13 |
| 2 | 2007-12-16 | 33000 | 배송완료 | 카드 | 2007-12-17 | 2007-12-20 |
| 3 | 2008-01-11 | 20000 | 배송완료 | 카드 | 2008-01-12 | 2008-01-13 |
| 4 | 2008-01-11 | 30000 | 배송완료 | 입금 | 2008-01-13 | 2008-01-15 |
| 5 | 2008-01-18 | 20000 | 주문취소 | 카드 | | |
| 6 | 2008-02-12 | 40000 | 배송중 | 카드 | 2008-02-13 | |
| 7 | 2008-02-20 | 20000 | 배송완료 | 카드 | 2008-02-21 | 2008-02-23 |

〈그림 3-19〉 통계정보를 집계하는 테이블 예시

- 업무규칙 : 월간 통계테이블의 수입액은 해당 년월에 해당되는 거래 완료된 주문의 총금액의 합과 비교하여 동일해야 한다.

〈표 3-40〉 계산 및 집계 관계 정합성 관련 업무규칙 SQL 예시

```
SELECT A.YEAR, A.MONTH, A.TOTAL_INCOME, B.SUM_INCOME
FROM   MONTHLY_STATISTICS A,
       ( SELECT SUM(ORDER_TOT_AMT) AS SUM_INCOME,
                TO_CHAR(REGDATE,'YYYYMM') AS YEAR_MONTH
         FROM   ORDER_DETAIL
         WHERE COMPLETE_DT IS NOT NULL
         GROUP BY TO_CHAR(REGDATE,'YYYYMM') ) B
WHERE A.YEAR || A.MONTH = B.YEAR_MONTH
  AND A.TOTAL_INCOME <> B.SUM_INCOME ;
```

## 제3절 데이터 품질측정

데이터 품질측정 단계에서는 사전에 도출한 업무규칙을 실제 운영 데이터베이스에 적용하여 오류데이터를 추출하고 오류율을 산출하여 오류현황을 파악한다. 데이터 품질을 측정함에 앞서 사전에 도출된 업무규칙을 토대로 데이터 측정규칙 목록을 작성한다. 측정규칙 목록에 업무규칙 ID · 해당 테이블명 · 컬럼명 · 업무규칙 정의 · 품질기준 · 업무규칙 SQL · 총건수 SQL · 오류건수 SQL 등을 명시한다. 〈표 3-41〉은 도출된 업무규칙 정의서를 토대로 업무규칙 SQL · 총건수 SQL · 오류건수 SQL을 도출하여 데이터 측정 규칙 목록을 작성한 예이다.

〈표 3-41〉 데이터 측정 규칙 정의 목록

| BR_ID | 테이블명 | 컬럼명 | 업무규칙 | 품질기준 | 규칙정의(SQL) | 총건수(SQL) | 오류건수(SQL) | 담당자 |
|---|---|---|---|---|---|---|---|---|
| BR_001 | TAB1 | COL1 | TAB1.COL1의 코드가 '2'인경우 COL2은 NOT NULL이어야 한다. | 조건완전성 | SELECT COL1, COL2 FROM TAB1 WHERE COL1='2' AND COL2 IS NULL | SELECT COUNT(*) FROM TAB1 | SELECT COUNT(COL1) FROM TAB1 WHERE COL1='2' AND COL2 IS NULL | 김규칙 |
| … | … | … | … | … | … | … | … | … |
| | | | | | | | | |

데이터 측정 규칙 정의서를 토대로 업무규칙을 실제 운영중인 물리 데이터베이스에 적용하여 해당 데이터 항목이 지켜야 할 규칙에 위반되는 오류데이터 발생 내역 · 오류 건수 · 진단 건수를 추출한다.

업무규칙을 적용한 데이터 품질측정은 샘플링을 실시하지 않고 운영 중인 데이터 레코드의 전수를 대상으로 측정해야 하므로 모든 업무규칙을 일괄 적용할 경우 운영시스템이 과부하를 줄 수 있다. 따라서 업무 외의 시간이나 시스템의 유휴 시간을 활용하여 데이터 품질을 측정하는 것이 바람직하다.

## 1. 업무규칙별 오류율 측정

사전에 준비한 측정 규칙 목록을 토대로 품질측정 SQL을 적용하여 총 건수 · 오류 건수 · 오류율을 계산한다. 해당 업무규칙의 테이블명 · 컬럼명 · 품질기준 · 업무규칙명 · 해당 테이블의 총 건수 · 오류 건수 · 오류율 등을 모두 포함하도록 취합한다. 〈표 3-42〉는 업무규칙별 오류율 측정을 위한 서식 구성 사례이며, 〈표 3-44〉에서 제시한 서식과 통합하여 하나의 품질측정 체크리스트를 만들어 사용하면 효과적이다.

〈표 3-42〉 업무규칙별 오류율 측정 예시

| BR_ID | 업무 | 영문 테이블명 | 한글 테이블명 | 영문 컬럼명 | 한글 컬럼명 | 품질 기준 | BR명 | 총건수 | 오류 건수 | 오류율(%) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 여신 | TAB1 | 계정 과목 | identifier | 계정 식별자 | 유일성 | 계정과목의 식별자는 유일해야 한다 | 36627 | 1 | 0.00 |
| 2 | 여신 | TAB1 | 계정 과목 | identifier | 계정 식별자 | 유효성 | 계정과목의 식별자의 데이터 형식은 '9999-99-9999'이어야 한다. | 36627 | 11 | 0.03 |
| 3 | 여신 | TAB1 | 계정 과목 | type_cd | 과목유형 | 일관성 | 계정과목의 과목유형은 통합코드데이블에 등록된 값을 사용해야 된다. | 36627 | 14 | 0.04 |
| 4 | 대출 | TAB2 | 거래 내역 | pcode | 거래코드 | 유효성 | 거래내역의 거래코드의 데이터 형식은 'SS_9999_9999_SS' 이어야 한다. | 18807 | 21 | 0.11 |
| 5 | 대출 | TAB2 | 거래 내역 | lnapp_dt | 대출 신청일 | 유효성 | 대출신청일은 'yyyy/mm/dd'의 날짜유형 이어야 한다. | 18807 | 30 | 0.16 |
| 6 | 대출 | TAB2 | 거래 내역 | lnapp_st | 상태 | 유효성 | 거래내역의 상태는 '01', '02', '03', '04', '05'의 값을 가진다. | 18807 | 12 | 0.64 |
| 7 | 상담 | TAB3 | 상담 접수 | ccode | 상담유형 | 유효성 | 상담접수의 상담유형은 '01', '02', 03'의 값을 가진다. | 336 | 13 | 3.87 |

## 2. 핵심 데이터별 오류율 측정

핵심 데이터 항목은 데이터 운영 조직의 경영과 고객 신뢰도에 매우 중요한 영향을 미치기 때문에 핵심정보항목과 매핑되는 업무규칙은 별도로 분류하여 품질을 측정한다.

〈표 3-43〉 핵심 데이터별 오류율 측정 예시

| 업무 | 핵심정보항목 | 컬럼명 | 전체건수 | 오류건수 | 업무규칙수 | 오류율 |
|---|---|---|---|---|---|---|
| 주문 | 상품코드 | PRD_CD | 4,123,441 | 10 | 3 | 0.000243 |
| | 운송장명 | TRNS_NM | 1,920,321 | 12 | 5 | 0.000625 |
| | 제조업체코드 | MKCP_CD | 1,023,021 | 234 | 2 | 0.022873 |
| | 원산지코드 | ORGN_CD | 1,023,021 | 12 | 6 | 0.001173 |
| | 출고구분 | DLV_CD | 2,012,121 | 23 | 3 | 0.001143 |
| | 배송구분 | TRNS_CD | 2,012,121 | 11 | 1 | 0.000547 |
| 고객 | 고객번호 | CUST_NO | 400,102 | 30 | 3 | 0.007498 |
| | 주민번호 | CUST_RSN | 400,102 | 29 | 4 | 0.007248 |
| | 고객주소 | CUST_ADR | 400,102 | 12 | 3 | 0.002999 |
| | 고객연락처 | CUST_TEL | 400,102 | 19 | 1 | 0.004749 |

취합된 데이터의 오류현황, 오류내역 등을 토대로 각 품질기준별 오류현황 등을 파악 한다. 또한 대상 시스템의 품질지수 등을 산출하고 품질측정결과 보고서를 작성하여 업무전문가와 업무 담당자에게 전달한다.

## 3. 데이터 품질 지수화

### 3.1 데이터 품질 지수

데이터 품질 지수는 품질측정을 실시한 대상 시스템의 데이터의 정확률을 대표하는 값을 의미한다. 일반적으로 정확률은 100에서 오류율(백분률)을 뺀 값으로 환산하여 100점 단위의 점수로 활용한다.

본 절에서는 해당 업무규칙별 중요도 값을 가중치로 선정하여 가중 평균(Weighted Average)을 계산하여 종합 오류율을 산출하는 방식을 예시한다. 가중치 선정 단위는 단일 업무규칙으로 선정하며 가중평균의 계산식은 다음과 같다.

$$\text{업무규칙 단위 오류율 } e_i = \frac{\text{오류 건수}}{\text{진단 데이터 항목의 총건수}}$$

$$\text{전체오류율(E)} = \frac{\sum_{i=1}^{n} w_i e_i}{\sum_{i=1}^{n} w_i} \qquad w_i\text{:업무규칙 중요도}$$

$$\text{전체품질지수(I)} = (1 - E) \cdot 100$$

다음은 업무규칙별 가중치를 적용하여 종합 품질 지수를 산출한 예이다.

〈표 3-44〉 종합 품질 지수 산출 예시

| 테이블 | 컬럼 | 업무규칙 | 품질기준 | 중요도 | | | | | | 총건수(N) | 오류건수(E) | 업무규칙별 오류율(E/N*100) | 가중 오류율 W·(E/N*100) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 테이블중요도 | 컬럼중요도 | 시스템연관성 | 서비스 영향도 | 업무 영향도 | 종합가중치(W) | | | | |
| TAB1 | COL1 | TAB.COL1은 TAB4.COL2에 등록된 값이어야 한다. | 일관성 | A | A | A | B | A | 2.8 | 13,021 | 10 | 0.08% | 0.22% |
| TAB2 | COL2 | TAB1.COL2는 유일한 값이어야 한다. | 유일성 | A | C | C | C | C | 1.4 | 32,101 | 21 | 0.07% | 0.09% |
| | COL3 | TABL2.COL3는 0이상 360이하의 값을 가져야 한다. | 유효성 | A | A | C | C | B | 2 | 32,101 | 13 | 0.04% | 0.08% |
| | COL4 | TAB2.COL4가 '04'이면 TAB4.COL3은 TAB5.COL4에 등록된 값이어야 한다. | 정확성 | A | A | A | A | A | 3 | 32,101 | 41 | 0.13% | 0.38% |
| TAB3 | COL5 | TAB3.COL5는 유일한 값이어야 한다. | 유일성 | B | C | C | C | C | 1.2 | 13,212 | 15 | 0.11% | 0.14% |
| TAB4 | COL6 | TAB4.COL6은 'YYYY-MM-DD'유형의 날짜 값을 가진다. | 유효성 | A | A | A | A | A | 3 | 14,321 | 231 | 1.61% | 4.84% |
| 가중평균 오류율 | | 0.43% | | | | | | | ΣW | ΣN | ΣE | ΣE/ΣN*100 | 가중 오류율합 |
| 품질지수 | | 99.57 | | | | | | | 13.4 | 136,857 | 331 | 0.24 | 5.75% |

산술 평균 오류율 = (0.08+0.07+0.04+0.13+0.11+1.61) / 6 = 0.34%
전체진단건수대비 오류율 = ΣE/ΣN*100 = 331 / 136,857 = 0.24%
가중 평균 오류율(E) = (0.08*2.8+1.4*0.07+2*.0.04+3*0.13+1.2*0.11+3*1.61)/(2.8+1.4+2+3+1.2+3) = 0.43%

가중치를 적용하였을 경우 가중평균 오류율은 0.43%, 가중치를 적용하지 않았을 경우 평균 오류율은 0.34%, 업무규칙별 전체오류건수를 전체 진단 건수로 나눈 전체 오류율은 0.24% 등으로 오류율 산정 방식에 따라 측정 결과가 서로 상이하다.
가중치를 적용하는 방식과 대표 값을 구하는 방식은 데이터 품질을 측정하는 기관에 따라 다르므로 데이터 품질 지수 및 오류율을 산출하기 전에 활용할 통계적 기법을 사전에 결정하여 확정한다.

## 3.2 중요도 산정

조직에서 운영하는 데이터는 관리 목적에 따라 평가되는 중요도가 서로 상이하므로 데이터 항목별로 중요도를 차등하여 관리할 필요가 있다. 예컨대 우편물을 취급하는 우체국이나 택배회사의 경우에는 고객의 주소가 매우 중요하나 텔레마케팅 회사에서는 조사에 활용할 고객의 전화번호가 더욱 중요하다. 중요도의 산정 및 평가 시기는 핵심 정보 항목을 도출하기 이전 단계 및 업무규칙 도출 이전 단계에서 수행된다. 평가된 중요도는 품질측정시 항목별 가중치를 차등 적용하기

위한 목적으로 활용된다.

〈표 3-45〉는 각 중요도 평가기준과 해당 항목에 대한 점수 부여기준을 나타낸 사례이다. 각 중요도별 점수부여 기준과 해당 중요도를 명시하고, 중요도를 채택 시 점수에 반영할 비율을 명시한다.

〈표 3-45〉 중요도 평가기준 선정 예시

| 중 요 도 | | 점수 부여 기준 | 중요도 | 채택 | 비율 |
|---|---|---|---|---|---|
| 테이블 중요도 | 현 시스템에서 차지하는 테이블 중요도 | 주요 마스터 | A | √ | 0.3 |
| | | 주요 결과/집계/이력 | B | | |
| | | 단순 참조 | C | | |
| 컬럼 중요도 | 현 시스템에서 차지하는 컬럼 중요도 | 식별자(UID) 속성 | A | | |
| | | 외부키 속성 | B | | |
| | | 추출 속성 | C | | |
| | | 단순 속성 | D | | |
| 시스템 연관성 | 여러 시스템간의 데이터 불일치시 발생 영향도 | 3개 이상 시스템 | A | | |
| | | 2개 시스템 | B | | |
| | | 1개 시스템 | C | | |
| 서비스 영향도 | 데이터 오류로 인해 장애를 유발할 수 있는 정도와 또는 고객의 불만사항을 제기하여 문제를 야기할 수 있는 정도 | 매우 중요 | A | √ | 0.7 |
| | | 중요 | B | | |
| | | 보통 | C | | |
| 업무 영향도 | 2개 이상의 서브시스템(고객관리시스템, 회계관리 시스템, 문서 관리 시스템)간의 업무상 연관 정도 | 3개 부서 이상 | A | | |
| | | 2개 부서 | B | | |
| | | 1개 부서 | C | | |

〈표 3-46〉은 업무 중요도와 서비스 중요도를 컬럼에 적용할 경우 중요도와 배점, 부여기준 등을 나열한 사례이다. 각각의 중요도는 아래의 사례에서와 같이 A, B, C와 같이 수준을 단계별로 구분하여 적용하며, 각 적용 중요도별 배점과 해당 중요도를 적용하는 기준을 명시한다.

〈표 3-46〉 테이블 중요도 세부 부여기준 예시

| 가중치 유형 | 중요도 | 배점 | 세부 부여 기준 |
|---|---|---|---|
| 테이블 중요도 | A | 3 | 기관 또는 조직의 근간이 되는 마스터성 핵심 데이터이며, 오류가 발생할 경우, 해당 업무 수행이 불가능한 컬럼 |
| | B | 2 | 마스터 데이터로부터 1차적으로 파생되는 상세 혹은 이력 정보 데이터 이며, 오류가 발생할 경우, 잘못된 정보를 제공하여 업무 수행에 지장을 줄수 있는 컬럼 |
| | C | 1 | 중요한 정보로 활용되는 기본 데이터나 참조성 데이터이며, 오류가 발생할 경우, 업무에 혼선을 줄 수 있는 잠재 컬럼 |

가중치를 계산하는 산식은 둘 이상일 경우 각각의 비중을 두어 계산을 하며, 가중치를 적용하는 산식은 〈표 3-47〉과 같다.

〈표 3-47〉 가중치 계산 및 적용 산식

| 최종 중요도 = (서비스 영향도 × 0.7) + (업무중요도 × 0.3) + ….. |
|---|

〈표 3-47〉는 컬럼별로 업무중요도와 서비스 중요도를 산정하여, 가중치를 계산한 예이다. 업무상 내부 이용자가 활용하는 데이터 보다 외부 고객이 이용하는 데이터에 대한 이용 만족도를 중요시하여 서비스 영향도를 0.7로 업무 중요도를 0.3으로 선정하여 컬럼 중요도를 평가한 사례이다. 일반적으로 가중치에 배점을 부여하는 산식은 가중치를 활용하는 조직에 따라 변경될 수 있다.

〈표 3-48〉 테이블 및 컬럼 중요도 목록

| 업무 | 영문 테이블명 | 한글 테이블명 | 영문 컬럼명 | 한글 컬럼명 | 데이터 타입 | 길이 | 중요도 | | | | | | 비고 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 테이블 중요도 | 컬럼 중요도 | 시스템 연관성 | 서비스 영향도 | 업무 영향도 | 중요도 | |
| 사업 관리 | PROJECT | 사업 | PRJ_NO | 사업번호 | VARCHAR2 | 8 | A | A | A | A | A | 3.0 | |
| | PROJECT | 사업 | PRJ_NM | 사업명 | VARCHAR2 | 255 | A | A | A | A | A | 3.0 | |
| | PROJECT | 사업 | DEPT_CD | 사업부처 | VARCHAR2 | 4 | A | A | A | A | A | 3.0 | |
| | PROJECT | 사업 | START_YY | 사업시작년도 | CHAR | 4 | B | A | B | B | B | 2.7 | |
| | PROJECT | 사업 | END_YY | 사업종료년도 | CHAR | 4 | B | A | B | B | B | 2.7 | |
| 간행물 관리 | PAPER | 논문 | KIND_GB | 학술지구분 | VARCHAR2 | 4 | A | A | A | A | A | 3.0 | |
| | PAPER | 논문 | ANNNC_DT | 논문발표일 | DATE | | A | B | A | A | A | 2.3 | |
| | PAPER | 논문 | TITLE | 제목 | VARCHAR2 | 255 | A | A | A | A | A | 3.0 | |
| | PAPER | 논문 | VOL_NM | 게게권/집 | VARCHAR2 | 128 | C | B | C | C | C | 1.7 | |
| | PAPER | 논문 | NUM_NO | 게재호 | NUMBER | 11 | C | B | C | C | C | 1.7 | |
| | PAPER | 논문 | START_PG_NO | 시작페이지 | NUMBER | 11 | C | C | C | C | C | 1.0 | |
| | PAPER | 논문 | END_PG_NO | 종료페이지 | NUMBER | 11 | C | C | C | C | C | 1.0 | |

# 제4절 오류 원인 분석

## 1. 오류 데이터의 발생 요인

정확한 데이터의 오류유형을 판정하기 위해선 사전 데이터에 대해 부정확한 값이 발생되고 유입되는 다양한 경로를 파악해야 된다. 일반적으로 부정확한 데이터가 발생하는 단계는 다음과 같은 유형으로 구분된다.

### 1.1 데이터의 입력 오류

데이터의 입력 과정에서 수기 입력을 허용하는 데이터 항목은 항상 입력자의 실수나 고의로 인한 오탈자 발생 가능성이 존재한다. 또한 정보시스템에서 필요 이상의 데이터를 필수 입력 값으로 요구할 경우 이용자가 입력을 기피하거나 꺼려

할 수 있어 e-메일 · 주소 · 휴대전화 · 체중 · 키 · 결혼여부 등과 같은 데이터 항목을 무의미하거나 거짓으로 입력할 수도 있다.

초기 정보 시스템을 구성하고 임의의 테스트 데이터를 입력한 이후, 시스템 정상가동 후에 테스트 데이터를 삭제하지 않은 경우 무의미하거나 부정확한 데이터가 입력될 수 있다. 일반적으로 많은 데이터베이스에서 이용자명에 '홍길동1', '테스트' 등의 데이터가 입력되어 있거나 전화번호 필드에 '111-111-1111', '999-999-9999' 등의 테스트 데이터가 지워지지 않고 잔재하는 경우를 흔히 볼 수 있다.

원천 문서를 입력하는 과정에서 과거 자필 문서 형식의 필체에 대한 인식이 불가하거나 고문서나 훼손되어 판독이 불가한 경우 입력자가 임의로 원천 문서를 해석하여 입력하는 경우를 흔히 볼 수 있다. 특히 한자로 된 고문서의 디지털화나 한자 문화권인 국내에서 한자 성명은 자필 문서일 경우 자획과 유사한 부수의 혼선으로 잘못된 정보를 입력할 수 있다.

또한 필수 입력 컬럼이 아닌 누락 값(NULL)의 의미 해석 과정에서도 문제점이 발생한다. 홍길동 영업사원의 휴대폰 번호가 누락된 상황은 '홍길동 사원은 휴대폰이 없다'와 '홍길동 사원의 휴대폰 번호를 모른다' 등 두 가지로 해석된다. 이러한 경우 입력자가 임의의 값이나 틀린 정보를 입력하도록 하지 않고 '알지 못함', '적용할 수 없음'의 유형을 구분하여 NULL을 설계하는 것이 바람직하다.

데이터의 입력 오류는 입력자의 실수 및 부적절한 입력 절차 외에도 시스템의 문제로 인하여 부정확한 값이 입력될 수 있다. 즉, 입력 시스템의 오류데이터 입력 통제 방식이 미비하거나 데이터베이스 내부의 무결성 제약조건을 방치하여 발생할 수도 있고, 입력 통제 방식은 적절하였으나 응용 프로그

램 및 시스템간 트랜잭션 처리 로직의 설계 미비로 인하여 누락된 값이나 중복된 값이 발생할 수도 있다.

### 1.2 시점 변화에 따른 데이터의 손상

초기에 정확했던 데이터가 시간이 지나면서 다른 시점에서는 부정확하게 될 수 있다. 데이터의 값은 항상 변경사항이 발생될 수 있으며 이는 데이터의 변경 시점 및 주기와 관련이 있다. 특정 시점에서는 정확한 데이터였지만 시간이 흘러 다른 시점에서는 부정확하게 된 데이터를 낙후데이터(Out-of-date Data)라 부른다. 낙후 데이터는 그 의미가 정확하지 못한 데이터에 해당되므로 부정확한 데이터의 범주에 속한다. 이러한 데이터의 사례는 개인의 이력정보, 물품의 재고상태, 계산 및 집계 값의 통계정보 등에서 흔히 찾아볼 수 있다. 예컨대 인사 데이터베이스의 경우 초기 입사 시점의 전화번호 · 주소 · 최종학력 등의 직원정보는 해당 직원의 연락처 변경 · 이사 · 재직 중 대학원 졸업 등의 사유로 이력 정보 변경이 발생될 수 있다. 이러한 정보 변경 내역을 시스템에 적절히 반영하지 못한 경우 현 시점의 데이터는 부정확한 데이터로 볼 수 있다.

반면 개인의 주민등록번호 · 생년월일 · 성별 등 개체의 내재적 속성은 시점의 변화와 무관하게 초기 데이터가 정확히 입력된 경우 지속적인 정확성을 유지할 수 있다. 따라서 데이터 항목의 변경 주기를 고려하여 데이터베이스를 설계할 필요가 있으며, 변경 관리가 필요한 데이터 항목은 적절한 재검증 시기와 주기를 사전에 정의하여 적절한 시기에 이력 변경사항을 적용하도록 유지하는 것이 바람직하다.

## 1.3 데이터의 흐름

단일 시스템의 유지보수 및 변경을 포함하여 다수의 정보 시스템과 연계가 있는 데이터베이스를 이동 및 변환 시 오류데이터의 발생 가능성이 높아진다. 특히 데이터웨어하우스나 데이터 마트를 활용하는 조직에서 데이터의 적절한 이동·변환·추출 규칙을 적용하지 못하거나 잘못된 데이터 정제 과정을 거칠 경우가 이에 해당된다.

데이터 이동 및 변환과정에서 부정확한 데이터의 발생 가능성을 줄이기 위해서는 원천시스템과 목표 시스템간의 불일치 문제를 우선 해결해야 된다. 따라서 데이터를 이동 및 변환시키는 원천 시스템과 변환된 데이터를 수용하는 목표 시스템의 값·구조·내용에 대해 정확히 이해해야 된다. 특히 현 시점에서의 원천 시스템의 데이터 현상을 파악하는 것이 매우 중요하다.

일반적으로 데이터의 이동 및 변환 과정은 ETL 도구를 활용하여 수행한다. ETL 도구는 추출·정제·변환·적재 과정을 지원한다. ETL 과정에서 발생할 수 있는 부정확한 데이터의 유형은 다음과 같다.

### 1.3.1 데이터 추출(Extract)

데이터 추출은 원천 데이터베이스로부터 다음 단계로의 이행 및 변환 작업을 위하여 원하는 데이터를 분리하여 뽑아내는 것이다. 데이터 추출 시 원천 시스템의 값의 현상 및 유효범위·정규화 수준·주요키·관계 정보 등을 사전에 파악하여 목표 시스템으로 데이터를 정확히 변환하기 위한 추출 로직을 작성해야 된다. 추출 로직이 잘못될 경우 정확한 변환·정제·적재 작업을 보장할 수 없다.

### 1.3.2 데이터 정제(Cleanse)

데이터 정제 과정은 오류데이터 값을 정확한 값으로 수정 혹은 삭제하는 것을 의미한다. 원천 시스템의 데이터를 잘못 분석하여 데이터 정제 규칙을 잘못 적용한 경우에는 오히려 데이터의 정확성이 떨어지게 된다. 원천시스템에 직원근무상태 코드가 '-999'란 코드가 사용되었으며 이 코드는 '임시 휴직'의 의미를 가지는 코드라고 가정하자. 만일 데이터 정제 담당자가 정제 규칙을 잘못 파악하여 비유효한 값으로 간주하여 해당 데이터 항목의 값을 NULL로 대체하거나 레코드를 삭제할 수 있다. 이러한 경우 과거데이터 또는 변환 전의 의미 있는 데이터가 소멸되게 되므로 향후 그 값을 되돌릴 방법이 없어진다.

### 1.3.3 변환(Transform)

변환 과정은 원천 시스템에 존재하는 값의 표현을 변경하여 목표 시스템의 데이터 항목의 데이터 규칙에 부합된 값으로 전환시키는 것을 의미한다. 원천 시스템에서는 여부 항목의 값을 '1', '0'을 사용하고 있고 목표 시스템에서는 여부 항목의 값을 'Y', 'N'으로 사용하고 있다면 여부 항목의 '1' 값을 'Y'로, '0' 값을 'N'으로 변환한다. 변환 과정에서도 원천 대 목표 시스템간의 변환 로직이 잘못되어 있다면, 부정확한 데이터를 발생시킨다.

### 1.3.4 적재(Load)

적재 과정은 추출된 데이터를 변환 및 정제하여 최종적으로 목표 시스템의 운영데이터로 적재하는 것을 의미한다. 적재 단계에서는 목표 시스템의 데이터 타입 · 길이 · 키정의 · 무결성 제약조건 등의 위반 때문에 적절히 적재를 못하고 거부된 데이터가 발생할 수 있다. 적재에 실패한 데이터는 오류 발생 원인을 분석하여 정확한 값으로 변환하여 다시 적재를 수행해야 된다.

신규 업무 시스템 개발 및 데이터 통합 작업은 개발 프로젝트의 마지막 단계에서 수행된다. 데이터 이행 작업은 정확한 추출 및 변환 로직으로 적재를 수행해야 되나 프로젝트의 종료 시기를 맞추기 위하여 적재에서 실패한 데이터를 처리하기 위해 적재되지 못한 데이터를 누락시키거나 목표시스템의 무결성 제약조건을 임의로 해제한 이후 로딩을 실시하는 과오를 범하는 경우가 빈번히 발생한다. 전자의 경우 중요 값의 누락이 발생될 수 있으며, 후자의 경우 무결성에 위배되는 데이터가 발생될 수 있다. 앞의 사례는 데이터의 적재 과정에서의 주요 오류발생 요인이 된다. 적재 작업 도중 실패한 데이터가 발생한다면 그 실패 요인을 명확히 분석하여 적절한 변환 및 정제 작업을 수행한 이후 데이터를 다시 적재해야 된다.

〈표 3-49〉 주요 오류데이터 발생원인

| 구분 | 설 명 | 오류 발생 원인 |
|---|---|---|
| 데이터 입력 | • 이용자가 데이터를 직접 입력<br>• 데이터를 수집하여 직접 생성한 사람이 아닌 다른 사람이 입력함 (입력 시점의 차이가 발생될 수 있음) | • 이용자 입력 오류<br>- 입력자 동기부여 수준 미비<br>- 입력자의 능력, 교육, 경험 수준<br>- 발생 시점과 입력시점 간의 차이<br>- 입력 전 자료이관<br>• 입력 통제 미비<br>- 누락 값 발생<br>- 중복 데이터 발생<br>- 유효성 체크 로직 미비<br>- 표준화 수준 미비(표준코드, 도메인)<br>- 주요 업무규칙 미처리<br>• 입력 절차의 문제점<br>• 입력 프로그램 오류<br>- 트랜잭션 처리 미비 |
| 데이터 흐름 | • 시스템 통합 이전의 과거데이터의 오류 잔재<br>• 주기적으로 타 시스템의 데이터를 수집하여 로딩을 수행<br>• 데이터 추출, 변환, 정제, 로딩과정에서 오류가 발생될 수 있음<br>• 분석 대상 데이터가 다른 데이터 원천에서 추출되어 생성된 경우, 추출 및 로딩 절차를 확인해야 함<br>• 원천 데이터와 목표 데이터베이스의 데이터 표준 및 구조를 정확히 파악해야 함 | • 데이터 변환 오류<br>- 표준화 오류 (코드, 허용값 오류 등)<br>• 데이터 정제 오류<br>- 부적합한 정제 로직으로 원천 데이터의 의미 상실<br>• 데이터 적재 오류<br>- 원천 데이터 오류<br>- 데이터 추출 및 변환 로직의 미흡<br>- 목표 시스템과의 부적합성<br>• 운영자 일괄 변경 오류<br>- 데이터 누락 또는 목표시스템 변경 |
| 데이터 재검증 활동 | • 주요 핵심 데이터에 대한 재검증 주기 및 변경 관리 절차가 수립되어야 함 (고객마스터의 주소, 연락처 등) | • 데이터의 재검증 절차 및 관리 프로세스 미흡<br>• 데이터의 최신성 확보를 위한 정책 미비<br>• 데이터 생성 및 수집 단계의 적시성 |

## 2. 오류 원인 분석 방법

업무규칙을 도출하여 데이터 품질을 측정한 이후 주요 오류 발생 컬럼에 대한 원인분석을 실시한다. 오류 원인 분석 시 선행되어야 할 작업은 오류 발생 컬럼의 오류발생 유형 · 오류발생 시점 · 오류발생 주기 · 오류가 발생된 데이터의 생성 및 변경 시점을 우선 파악한다.

파악된 오류 발생 유형 · 오류 발생 시점 · 데이터 생성 시점을 토대로 과거 데이터 통합 이전의 데이터, 입수된 데이터 혹은 지속적으로 발생되는 데이터인지를 파악하여 오류데이터 발생 추이를 분석한다. 또한 데이터의 오류발생 유형에 대해서도 데이터 입력 통제 · 데이터 흐름 · 과거 미보정 데이터인지를 판정한다.

### 2.1 데이터의 입력 통제

데이터 입력 통제는 입력자가 응용프로그램에서 데이터를 입력할 때 부정확한 데이터를 탐지하여 입력 절차에 제한을 주는 것을 의미하며, 입력자가 부정확한 데이터를 입력하는 실수를 미연에 방지하는 역할을 한다. 응용프로그램 설계 및 구현 시 부적절한 입력 통제 절차로 인하여 다양한 유형의 부정확한 데이터를 유발시킨다. 앞 절에서 설명하였듯이 데이터의 입력 오류는 입력자의 실수 및 고의적 오류데이터 생성, 입력 절차 및 시점의 문제점 등 매우 다양한 오류데이터 유발 요소를 내포하고 있으므로 적절한 데이터 입력통제 정책을 수립하여 응용프로그램에 반영해야 된다.

오류가 발생된 업무규칙의 주요 원인이 응용프로그램일 경우 오류가 발생된 주요 응용프로그램과 관련 모듈을 추적해야 된다. 오류데이터가 발생한 응용프로그램을 조사하기 위해서 오류가 발생된 DB-응용프로그램 · Table-응용프로그램 · DB

Object 간의 CRUD Matrix 상관성 분석을 실시하여 해당 프로그램 모듈을 추적하고 결함 원인을 분석한다.
〈표 3-50〉은 다음의 업무규칙 및 추정 오류 사례에서 CRUD Matrix를 활용한 응용프로그램간의 상관 분석을 실시하여 주요 오류발생 애플리케이션을 추적하는 예제이다.

- 업무규칙 : 구매주문항목의 구매주문일련번호 값이 존재하고 주문구매항목의 구매주문항목 구분코드가 존재하면 구매주문항목 입력일은 반드시 존재해야 된다.
- 측정결과 : 총 30,120 레코드 중에서 오류데이터 102건 발생
- 오류발생 시점 및 주기 : 구매주문항목 입력 시 지속적으로 발생하며,
  트랜잭션이 잦은 오전 10시부터 12사이에 자주 발생
- 추정원인 : 입력 통제 오류 (업무시스템 업그레이드 시점 이후로 간헐적으로 오류데이터 발생, 트랜잭션 처리 로직 오류로 추정)

실제 레코드가 생성되는 정보항목은 구매 주문항목 테이블이며, 관련 프로세스는 구매 주문 항목 추가 프로세스이므로, 구매 주문 항목 추가 응용프로그램 모듈을 추적하여 오류발생 원인을 파악할 수 있다.
데이터 입력 통제는 응용프로그램 뿐만 아니라, 데이터베이스 운영 시스템에 적용 가능한 제약 조건 설계에도 적용하여 오류데이터가 데이터베이스에 입력되는 것을 미연에 방지해야 된다.

〈표 3-50〉 테이블/애플리케이션 상관분석의 예시

| 공백 = 해당없음<br>C = 생성(Create)<br>R = 참조(Read)<br>D = 삭제(Delete)<br>U = 수정(Update)<br>기본프로세스 | 정보 항목 | 고객 | 제품 | 창고 | 재고 항목 | 공급자 | 구매 주문 | 구매 주문 항목 | 판매 주문 |
|---|---|---|---|---|---|---|---|---|---|
| 신규고객등록 | | C | | | | | | | |
| 구매주문생성 | | | R | | | R | C | C | |
| 구매주문항목추가 | | | R | | | R | R | C | |
| 재고항목 조사 | | | | R | U | | | | |
| 판매주문생성 | | R | R | | | | | | C |
| … | | … | … | … | … | … | … | … | … |

## 2.2 데이터 흐름 통제

데이터 흐름 통제는 원천 데이터를 수기로 생성하거나 추출 · 변환 · 적재를 통해 생성하여 목표 시스템으로 이동시 데이터 흐름 규칙을 적용하여 부정확한 데이터가 입수되지 못하도록 통제하는 것을 의미한다. 데이터 흐름 통제를 적절히 수행하지 못하는 경우 입력 통제와 유사하게 부정확한 데이터가 발생될 수 있다.

업무규칙 측정 결과 오류데이터 발생 원인이 데이터 흐름과 관련된 경우 소스 대 타깃 매핑 분석과 소스 데이터 추출 모니터링 보고서 등의 검토를 통하여 그 원인이 흐름과정에서 추출 · 변환 · 적재 조건의 오류에 기인한 것인지 검토한다.

〈표 3-51〉는 아래의 업무규칙 및 추정 오류 원인에 대하여 소스 대 타깃 매핑 분석을 통하여 주요 오류발생 원인을 분석하는 예제이다.

- 업무규칙 : 원장 마스터(TG_003)의 상태코드(STAT_CD)는 '01', '02', '03'의 값을 가져야 한다.
- 측정결과 : 총 310,201 레코드 중에서 오류데이터 123건 발생 ( '00', '09', '0N' 등의 미정의 코드 존재)
- 오류발생 시점 및 주기 : 상시 발생, 원천 데이터 추출 시 발생
- 추정원인 : 원천 데이터 변환 로직 오류로 추정

위의 업무규칙을 적용한 결과 미사용 코드인 '00', '09', '0N' 등의 오류데이터가 발견되었다. 데이터 이행 시 적용한 소스 대 타깃 매핑 분석서를 검토하였을 경우 원천 데이터를 목표 시스템으로 변환 시 과거 오류데이터의 변환 로직을 고려하지 않고 단순히 코드 앞에 '0'을 추가하여 자릿수만 일치시키는 변환 로직을 사용한 것을 알 수 있다. 이러한 경우 원천 데이터의 '0', '9', 'N'의 처리 값을 규명하여 적절한 변환 로직 또는 정제 작업 수행 후 재차 적재 작업을 수행해야 된다.

〈표 3-51〉 소스 대 타깃 매핑 분석서 예시

| 원천 | | | | 목표시스템 | | | | 변환로직 |
|---|---|---|---|---|---|---|---|---|
| 테이블명 | 컬럼명 | 데이터 타입 | 길이 | 테이블명 | 컬럼명 | 데이터 타입 | 길이 | |
| ST_001 | CUST_CD | CHAR | 4 | TG_001 | CUST_CD | CHAR | 8 | 코드 앞에 '0000' 추가 |
| ST_002 | APPL_NO | CHAR | 8 | TG_002 | APP_NO | CHAR | 12 | 신규 승인 신청번호로 변경 |
| ST_002 | GWA_CD | CHAR | 4 | TG_002 | PRD_CD | CHAR | 8 | 신규 상품 코드로 변경 |
| ST_003 | STAT_CD | CHAR | 1 | TG_003 | STAT_CD | CHAR | 2 | 기존 코드 앞에 '0' 추가 |

## 2.3 과거 미정제 데이터

단일성 데이터 통합 및 신규 응용프로그램 개발, 시스템 업그레이드 등의 프로젝트에서 데이터 이행 작업은 프로젝트의 최종 단계에서 수행된다. 프로젝트 종료 후 시스템이 가동되면 추가 정제 및 변환 작업을 수행하기 어렵다. 만일 데이터 이행 작업 시 적절한 데이터 변환 작업을 수행하지 못한 시스템은 차기 데이터 정제 작업 전까지 지속적으로 오류데이터를 제공할 수밖에 없다. 일부 응용프로그램 또는 시스템은 과거 오류데이터가 서비스 되지 않도록 강제화하여 이용자들은 오류데이터가 시스템 내부에 은닉되어 있는지 파악하기 어려울 수 있다.

과거 미정제 데이터는 오류가 발생되는 레코드와 해당 레코드의 생성 및 갱신 시점· 과거 시스템 변경 시기·최근 데이터 통합 또는 데이터 이행 시점 등을 파악하여 오류 원인 분석을 수행한다.

- 업무규칙 : 논문의 심의여부코드는 '1:통과', '2:미통과', '3:보류'의 값을 가져야 한다.
- 측정결과 : 총 13,213 레코드 중에서 오류데이터 121건 발생 ( '1', '2', '3' 외의 'N', '0' 등의 미정의 코드 존재)
- 오류발생 시점 및 주기 : 상시 발생
- 추정원인 : 과거 미정제 오류데이터로, 현 입력시스템에서는 적절히 반영 되었으나, 2003년 시스템 업그레이드 시점 이전에 생성된 데이터임

데이터베이스 시스템은 현 IT 상황과 비즈니스의 요구사항에 따라 지속적으로 업그레이드·통합·데이터 이행 작업이 반복

된다. 이러한 반복과정에서 정제되지 못한 오류데이터는 시스템의 변경 과정에서 누락되거나 정제되었더라도 원 의미를 상실할 수 있으며, 시스템의 업그레이드와 데이터 이행작업이 수년간 지속될 경우 과거 오류데이터는 지속적으로 발생될 수 있다. 따라서 오류발생 원인이 과거 미정제 데이터일 경우 적절한 정제 작업을 수행하여 향후 시스템 업그레이드 시 동일한 오류 데이터가 발생되지 않도록 조치해야 된다.

## 3. 품질기준별 원인분석 사례

### 3.1 누락된 값 (NULL)

누락된 값은 크게 물리적 의미의 누락된 값과 업무적 의미의 누락된 값으로 구분할 수 있다. 누락 값의 점검은 NULL 값 및 공백 값의 존재 여부, ID 성격의 컬럼에 Null 존재 여부, 업무상 공통 필수 입력 항목 등을 점검하는 것을 포함한다.
Null 데이터 입력 시 필수 입력 항목일 경우 데이터의 입력 통제가 적절히 이루어진다면 물리적 의미의 누락 값은 피할 수 있다. 또한 물리 데이터베이스에 NOT NULL이나 DEFAULT 등의 제약 사항을 적용하면 응용프로그램에서 미처 처리 못한 NULL을 사전에 예방할 수 있다. 다만 업무적 의미의 누락 값은 다양한 업무 프로세스의 요구사항에 따라 발생될 수 있으므로 초기 데이터 입력 통제 절차에서 공통 필수 입력 항목 또는 각 입력 데이터 항목별 다수 필수 입력 조건을 적용하여 입력 이벤트에 반영함으로써 누락 값의 오류를 최소화할 수 있다.

- 업무규칙 : 특허마스터의 특허 출원일이 존재하면 특허 출원번호는 반드시 존재해야 된다.
- 측정결과 : 총 340,124 레코드 중에서 오류데이터 316건 발생

- 오류발생 시점 및 주기 : 상시 발생
- 추정원인 : 입력통제 로직 미비. 출원일 입력 시 연관된 데이터 항목인 출원번호를 반드시 입력해야 하나, 입력 통제 미비

위의 사례는 특허가 출원되었으면 특허 출원일을 입력하고, 출원일이 입력되었을 경우 출원번호는 동시에 입력해야할 필수 항목이나 입력 통제의 미비로 인하여 출원번호가 입력되지 못한 사례이다.

## 3.2 유효성 위반

유효성은 정해진 데이터 유효 범위를 충족하고 있는 데이터를 의미한다. 유효범위는 특정 개체가 가지는 속성 값이 가질 수 있는 범위라 할 수 있다. 유효성은 데이터 모델링 단계에서 정의되며, 속성의 자료형 · 길이 · 정밀도 · 유효범위 등으로 표현된다. 그 외에도 날짜유형 · 주민등록번호 · 사업자등록번호 등의 자료형 및 길이와 같이 일반적 값의 범위 외에 실제로 정확해야 할 데이터 도메인이 존재한다. 날짜유형의 경우 날짜의 말일이 존재하며, '20070231'은 비유효한 날짜로 인식된다. 또한 '123456-1234567'과 같은 주민등록번호는 실제로 존재하지 않는 주민등록번호의 유효범위로 인식된다. 본 절에서는 값의 다양한 유효 범위를 분류하여 '값의 범위 유효성', '허용값 목록 유효성', '날짜 유효성', '포맷 유효성', '특수 도메인 유효성'으로 구분한다.

### 3.2.1 값의 범위 유효성

값의 범위 유효성은 속성이 가질 수 있는 값의 일반적인 범위 또는 업무적으로 지정된 범위를 충족해야 하는 경우 적용한다. 일반적으로 각도의 값은 0에서 360도 미만의 값을 가져야 하나 -9 등의 값이 입력된 경우나, 연령의 경우 0 이상의 값을 가져야하나 0 미만의 값을 가질 경우 값의 범위 유효성에 위배된 오류데이터로 볼 수 있다.

- 업무규칙 : 환자의 맥박수는 0 이상이어야 하고 300 미만이어야 한다.
- 측정결과 : 총 123,111 레코드 중에서 오류데이터 45건 발생. 0, 10, 1555, 1200, 800, 11100 등의 일반적 수치 이상 및 미만의 값 다수 존재
- 오류발생 시점 및 주기 : 상시 발생
- 추정원인 : 입력통제 로직 미비. 이용자의 오자 입력으로 인해 기본 값의 범위는 충족하였으나 1555, 1200, 800 등의 엉뚱한 값에 대한 입력 통제 미비

위의 사례는 입력자가 시스템에 데이터 입력 시 키보드를 잘못 입력하여 일반적 값의 허용범위를 초과하는 데이터를 입력한 것으로 볼 수 있다. 입력 값의 통제 시 적정수준의 최대값과 최소값을 유지하여 입력통제 로직을 반영함으로써 입력자로 하여금 실수를 최소화 하도록 하는 것이 바람직하다.

### 3.2.2 허용값 목록 유효성

값의 허용값 목록 유효성은 속성이 가질 수 있는 값의 일반

적인 집합, 즉 유효값의 목록을 충족해야 하는 경우 적용한다. 일반적으로 성별('M', 'F') 컬럼이나, 여부('Y', 'N') 컬럼 등과 같이 코드성 컬럼을 별도로 정의하여 사용하는 경우 허용값 목록 유효성을 적용한다. 허용값 목록의 유효성 오류는 개발자가 코드 표준을 참조하지 않고 임의로 별도 코드를 작성해 응용프로그램을 개발하거나 수기로 코드를 직접 입력할 수 있도록 응용프로그램에서 허용하는 경우 빈번히 발생한다.

- 업무규칙 : 지적재산권의 상태는 '01' 심의중, '02' 출원, '03' 등록, '04' 의견제출, '05' 거절, '06' 포기 의 값을 사용해야 된다.
- 측정결과 : 총 123,211 레코드 중에서 오류데이터 15건 발생. '출원', '등록', '드록', 'cnfdnjs', '-99' 등의 오류데이터 존재
- 오류발생 시점 및 주기 : 2006년도 이전데이터에서 일부 발생
- 추정원인 : 과거 시스템의 입력통제 로직 미비로 인한 과거 오류데이터의 존재. 이용자의 오자 입력으로 인해 '출원', '등록', '드록', 'cnfdnjs', '-99' 등의 오류데이터 존재

위의 사례는 신규시스템 개발 이전의 과거 시스템에 수기 입력된 오류데이터가 현행화 되지 못하여 현재까지 발생되는 오류 사례이다. 현재는 입력 시스템의 업그레이드로 오류데이터가 발생하지 않으나 과거 오류데이터는 사용된 코드 값을 명확히 구분하여 정제작업을 수행해야 한다.

### 3.2.3 날짜 유효성

날짜 유효성은 해당 속성이 주어진 날짜 유형을 준수해야 함을 의미한다. 날짜 유효성 위반 사례는 날짜유형을 데이터베

이스 내장 데이터 타입인 DateTime 유형을 사용하지 않고, CHAR 또는 VARCHAR 유형의 문자형을 날짜유형으로 사용할 경우 주로 발생된다. 일반적으로 '2008-02-31', '2008-06-31'과 같이 말일을 31로 임의 입력하거나 'YYYY-MM-DD' 유형과 'YYYYMMYY' 유형의 날짜를 동일 속성에 일관되지 않게 입력하는 경우가 주요 날짜 유효성 위반 사례이다.

- 업무규칙 : 특허마스터의 출원일은 'YYYY-MM-DD'의 날짜유형을 준수해야 된다.
- 측정결과 : 총 131,821 레코드 중에서 오류데이터 1,215건 발생. 'YYYYMMDD' 유형 901건(2000년도 이전 과거 데이터), 비유효날짜 '1972-02-31', '1934-06-31' 등의 데이터 314건 발생(최근 데이터)
- 오류발생 시점 및 주기 : 상시발생
- 추정원인 : 입력통제 로직 미비. 과거 응용프로그램에서는 수기 입력을 허용하였으나 2000년도 이후 신규 응용프로그램에서는 날짜 자리수의 기본검증은 실시하였음에도 불구하고 실제 유효한 날짜의 검증로직을 적용하지 못함

위의 사례는 응용프로그램에서 날짜의 포맷 유효성만 확인하고 실제 존재하는 날짜여부를 확인하지 못하는 입력 통제오류의 전형적인 사례이다.

### 3.2.4 포맷 유효성

포맷 유효성은 속성 값이 지켜야할 데이터 형식 및 자리수를 준수해야할 경우 적용한다. 직원의 휴대폰 번호는 999-9999-9999의 포맷 또는 999-999-9999 (C:영문자, 9:숫자)을

준수해야 하나 유효 포맷을 검증하는 절차가 누락된 경우 포맷 유효성에 위반된 데이터가 발생할 수 있다.

- 업무규칙 : 자산마스터의 자산코드는 CC-9999-C999의 포맷을 준수해야 된다.
- 측정결과 : 총 11,098 레코드 중에서 오류데이터 78건 발생, C-9999-C999, C-999999-C999, 999999-C999 등의 오류 데이터 유형 다수 발생
- 오류발생 시점 및 주기 : 상시발생
- 추정원인 : 입력통제 로직 미비 및 응용프로그램 자동 코드 생성 오류

위의 사례는 응용프로그램의 자산코드를 자동 생성하는 모듈에서 앞의 신청부서코드-신청년월-자산유형코드-일련번호 형태의 자동 생성된 코드를 임의 부여하도록 되어 있는데, 부서코드가 누락되었거나 신청년도만 존재할 경우 시스템에서 알고 있는 정보만의 조합으로 코드를 생성하도록 작성되어 비유효한 포맷의 자산 코드가 생성되는 예제이다. 비유효한 포맷으로 생성된 코드는 임의 저장이 되지 않고 입력화면에서 이용자가 입력 확인 후 저장 되도록 되어 있으나 입력 통제 시 유효 포맷을 확인하지 않고 누락 값 체크만 하도록 되어 오류 데이터가 입력된 사례이다.

### 3.2.5 특수 도메인 유효성 위반

특수 도메인은 주민등록번호 · 사업자등록번호 · ISBN · ISSN 등과 같이 번호 자리수의 조합 또는 번호의 특정 자리수에 체크 디지트(Check Digit)가 존재하여 사실데이터 확인을 위해 별도의 유효성 검증이 필요한 도메인을 의미한다. 이러한 도

메인의 유효성은 별도의 검증 함수 또는 프로시저를 작성하여 데이터의 유효성 검증을 실시한다.

- 업무규칙 : 고객의 주민등록번호는 주민등록번호 형식을 준수해야 된다.
- 측정결과 : 총 45,123 레코드 중에서 오류데이터 128건 발생

  '123456-111111', '121231-1234567', '781210-0000000' 등의 주민등록번호 유효성 검증 불일치 데이터 다수 발생
- 오류발생 시점 및 주기 : 상시발생
- 추정원인 : 입력통제 로직 미비

위의 사례는 데이터 입력자가 입력하는 데이터 포맷만을 검증하고, 실제 유효데이터  및 비유효 데이터를 검증하는 함수를 적용하지 않는 경우 발생되는 사례이다. 특수 도메인의 유효성 검증은 위의 사례와 같이 별도의 검증 함수를 적용하여 실시한다.

### 3.3 유일성 위반

유일성은 실세계에서 하나 이상 존재할 수 없는 속성 또는 업무적으로 유일하게 식별되어야 하는 속성의 중복이 발생하지 않음을 의미한다. 일반적으로 단일 속성의 유일성은 마스터성 테이블의 식별자성 속성에 적용하며, 마스터 테이블 이외의 이력 또는 거래내역 등의 테이블은 복수 컬럼의 조합으로 중복값을 발견할 수 있다. 유일성 품질점검 기준은 '단일 컬럼 유일성', '복합 컬럼 유일성'으로 구분된다.

### 3.3.1 단일 컬럼 유일성

단일 컬럼 유일성은 마스터성 테이블의 ID성 컬럼의 중복 값이 없어야 할 경우 적용한다. 유일성 측정의 대상이 되는 속성은 마스터 테이블의 주민등록번호 · 사업자등록번호 · 법인등록번호 · 특허번호 등의 식별자성 컬럼이 이에 해당된다.

- 업무규칙 : 기관마스터의 사업자등록번호는 유일해야 된다.
- 측정결과 : 총 11,023 레코드 중에서 오류데이터 12건 발생
- 오류발생 시점 및 주기 : 상시발생
- 추정원인 : 입력통제 로직 미비

위의 사례는 데이터 입력 시 중복데이터 검증 로직을 적용하지 않고 값의 포맷이나 누락 여부만 파악하는 기본 검증 로직을 적용하여 마스터성 테이블에 중복된 레코드가 입력되는 사례이다. 운영 DBMS 설계 시 유일성 제약조건(Uniqueness)을 적용하면 응용프로그램에서 적용하지 못한 수작업 입력에 따른 값의 중복을 미연에 방지할 수 있다.

### 3.3.2 복합 컬럼 유일성

복합컬럼 유일성은 둘 이상의 컬럼의 조합으로 유일해야 하는 경우 적용한다. 일반적으로 거래, 이력 등의 트랜잭션 테이블의 중복된 레코드를 선별하거나 업무적으로 특정 시점에서 동일 장소와 시간에 특정 주체가 동시에 발생할 수 없는 데이터를 검증하고자 할 때 적용한다. 대표적인 사례로는 도서의 대출 · 회의실의 예약 · 자동차의 임대 등이 해당된다.

- 업무규칙 : 차량번호, 임대자 ID, 임대시작일을 조합하였을 경우 유일해야 된다.
- 측정결과 : 총 5,903 레코드 중에서 오류데이터 42건 발생
- 오류발생 시점 및 주기 : 상시발생
- 추정원인 : 입력통제 로직 미비

위의 사례는 응용프로그램에서 복합된 키의 조합으로 인한 중복데이터 검증로직을 적용하지 않아서 이력 및 트랜잭션 테이블에 중복된 데이터가 발생한 사례이다. 운영 DBMS 설계 시 유일성 제약조건(Uniqueness)을 다수의 키컬럼으로 적용하면 복합 컬럼의 중복을 방지할 수 있다.

## 3.4 일관성

일관성은 정보시스템 내의 동일한 데이터 간에 불일치가 발생하지 않음을 의미한다. 일관성 오류는 데이터에 대한 불분명한 정의, 데이터 참조 무결성(Referential Integrity) 결여 혹은 개별 시스템 단위로 설계되어 전사 관점의 코드 통합 수준이 미흡한 경우 주로 발생한다. 일관성은 '참조 무결성'과 '코드 일관성'으로 구분된다.

### 3.4.1 참조 무결성

참조 무결성은 데이터 모델에서 정의된 객체 간의 관계 조건을 유지해야한다는 의미를 가지고 있으며, 둘 이상의 테이블 및 컬럼 간에 상호 연관 관계가 있는 경우 해당 테이블과 테이블 혹은 컬럼과 컬럼 간의 동일한 값을 유지해야 할 경우 적용한다.

- 업무규칙 : 논문의 과제ID는 과제기본정보의 과제ID의 값에 반드시 존재해야 한다.
- 측정결과 : 총 12,612 레코드 중에서 오류데이터 153건 발생
- 오류발생 시점 및 주기 : 상시발생
- 추정원인 : DBMS 참조 무결성 제약조건 결여 및 데이터 모델 변경관리 정책 미흡

위의 사례는 해당 과제기본정보를 삭제하였지만, 과제기본정보와 참조관계에 있는 논문 레코드를 삭제하지 않아서 과제ID가 남아 있는 것으로 볼 수 있다. 참조 무결성 오류의 발생을 방지하려면 해당 데이터의 입력·수정·삭제 시 참조관계에 있는 데이터가 어떻게 처리되어야 할지를 사전에 정의해야 되며, 물리모델링에서 확정된 구조정보를 실제 운영 데이터베이스에 참조 무결성 제약조건으로 적용해야 된다. 또한 데이터 모델의 변경관리 정책을 수립하여 데이터 모델-DB간 최신성을 유지해야 된다.

### 3.4.2 코드 일관성

전사적 관점에서 표준코드·도메인·유효 데이터 제약사항 등의 사전 분석이 선행된 이후 시스템을 개발해야 되나, 전사 표준화 코드·도메인 분석을 수행하지 않고 응용프로그램 개발에 치중함으로써 개발자 나름대로의 입력 코드를 작성하거나 비표준 코드 입력, 개별 코드를 필요시마다 만들어 입력하게 되는 경우 등이 발생하여, 데이터의 일관성을 해치는 무분별한 데이터 값이 생성되어 활용되는 오류가 발생할 수 있다. 예컨대 통합코드 마스터 테이블에는 '민원상태코드'가 '1:접수', '2:처리', '3:완료', '4:보류'로 정의되어 있는데, 실제 사용된 민원처리 테이블에서는 '1', '2', '3', '4', '5', '6'의 값이 사

용'되었을 경우 표준코드에 미등재된 '5', '6' 값을 사용하므로 코드 일관성이 결여된 오류데이터에 해당된다. 이러한 경우 코드 일관성을 적용한다.

- 업무규칙 : 고객원장의 고객상태는 표준코드테이블의 구분코드 '010'(고객상태) 에 등재된 값을 사용해야 된다.
- 측정결과 : 총 1,020,165 레코드 중에서 오류데이터 1,131건 발생
  '11', '112', '113', '??', '-99', '--' 등의 데이터 다수 존재
- 오류발생 시점 및 주기 : 상시 발생
- 추정원인 : 데이터 입력 통제 오류 · 수기 입력에 따른 데이터 오류 · 데이터 표준관리 미흡

고객상태 컬럼은 준수해야 하는 표준코드가 정의되어 있고 이에 따라 표준코드에 등재된 값을 사용해야 되나 위 사례에서는 표준코드에 미등재된 임의 입력 값으로 보이는 '11', '112', '113', '--' 등의 오류데이터가 입력된 것으로 보인다.

### 3.4.3 정확성

정확성 품질점검 기준은 실세계에 존재하는 객체의 표현 값이 정확히 반영되어야 함을 의미한다. 정확성 품질기준은 그 의미가 가장 광범위하며, 앞서 언급한 품질기준 이외의 데이터 규칙과 프로세스 규칙을 포괄하는 다양한 업무규칙을 점검하기 위한 품질기준이다.

- 업무규칙 : 고객 등급 '골드'인 회원은 해당 년도에 기산연월 2007년 3월부터의 실적 포인트가 300 이상이거나 연평균 잔액 실적이 5,000만 원 이상이어야 한다.

- 측정결과 : 총 120,111 레코드 중에서 오류데이터 116건 발생
- 오류발생 시점 및 주기 : 상시 발생
- 추정원인 : 입력통제 미비. 고객 등급 입력 및 수정 절차에서 고객 등급을 판정하는 주요 업무규칙 누락. 입력자 임의의 고객 등급 수정 가능

위의 사례는 고객의 등급 입력 시 특정 기간과 고객 실적 등의 계산 규칙에 따라 자동 계산된 고객의 등급을 입력 수정하게 되어 있으나 고객등급 값의 변경이 허용되어 오류데이터가 발생된 사례이다. 다수 컬럼의 계산 값 및 업무로직을 적용하여 정확성이 확보되어야 하는 데이터 항목은 별도의 업무규칙 검증 로직을 사전에 마련하여 입력 통제를 실시해야 된다. 측정 주기와 시점이 명시된 업무규칙은 지속적 관리 및 주기적 모니터링이 필요하다.

# 제4장 데이터 품질진단 기법 - 비정형 데이터

본 장에서는 비정형 콘텐츠에 적용할 수 있는 품질진단 기법에 대해 기술한다. 비정형 콘텐츠의 품질측정에 대해 사용할 체크리스트의 구성과 작성에 대해 기술하고, 체크리스트에 따른 품질측정 결과로부터 품질지수를 산출하는 방법에 대해 기술한다.

## 제1절 데이터 프로파일링 및 업무규칙 도출

비정형 데이터 즉, 디지털화된 멀티미디어 데이터는 동영상, 이미지, 사운드 등과 같은 비정형 콘텐츠 자체와 비정형 콘텐츠의 메타데이터로 관리 포인트를 나누어 고려해야 한다.
비정형 콘텐츠 자체는 특정한 포맷의 파일 형태로 작성되어 데이터베이스 내부나 외부에 저장되어 활용되고, 이들의 메타데이터는 데이터베이스 내에 별도의 저장 구조로 설계되어 관리된다. 비정형 콘텐츠에 대한 메타데이터는 크게 콘텐츠

자체에 내장된 메타데이터와 데이터베이스 내에 별도 설계를 통해 정형 텍스트 데이터 형태로 관리되는 메타데이터로 구분해 볼 수 있다. 이들 중 데이터베이스 내에 별도의 정형 텍스트 데이터 형태로 존재하는 메타데이터에 대해서 데이터 프로파일링 및 업무규칙의 적용이 가능하며, 3장의 정형 텍스트 데이터에 대한 데이터 프로파일링 및 업무규칙 도출에서 기술한 내용을 동일하게 적용할 수 있다.

비정형 콘텐츠의 메타데이터에 대한 프로파일링 및 업무규칙 도출을 위해 비정형 콘텐츠의 메타데이터 구성 항목에 대한 평가와 이들 간에 존재하는 업무규칙을 도출하기 위한 관련 절차가 수행되어야 한다.

## 제2절 체크리스트 준비

운영 중인 비정형 콘텐츠에 대한 품질측정을 위해서는 콘텐츠 유형에 따른 품질기준이 우선 정의되어야 한다. 콘텐츠 유형별로 적용 가능한 품질기준은 서로 상이할 수 있으며, 품질측정 시 해당 콘텐츠 유형에 따른 모든 품질기준에 대해 측정을 할 수도 있고, 일부 품질기준에 대해 측정을 할 수도 있다. 즉, 비정형 콘텐츠에 대한 품질측정은 품질측정의 대상 및 범위에 대해서 먼저 고려가 이루어지고 난 후 그에 따른 품질측정 항목 구성 및 측정 내용 구성이 수행되어야 한다. 비정형 콘텐츠에 대한 품질측정을 위한 측정 기준 선택과 이에 따른 측정 항목 및 측정 내용은 체크리스트로 정리되고 품질측정 시 이 체크리스트에 수록된 측정 내용에 대해 점검함으로써 품질측정에 직접적으로 활용된다.

체크리스트는 비정형 콘텐츠에 대한 품질측정 시마다 준비되

어야 하는데, 선정된 측정 기준과 이에 따른 측정 항목 및 측정 내용을 측정 시마다 필요에 따라 임의로 작성하여 체크리스트를 구성하거나 측정 기준과 측정 항목, 측정 내용을 리파지토리화하여 선정할 수도 있다. 리파지토리를 이용하는 방법은 다양하고 수많은 측정 항목과 측정 내용을 효과적으로 관리하면서 과거에 선정된 구성을 재활용하거나 재구성하는데 매우 용이하기 때문에 적극 권장할 만하다.

## 1. 측정 기준의 선정

〈표 2-5〉, 〈표 2-6〉, 〈표 2-7〉, 〈표 2-8〉에서 제시한 비정형 콘텐츠의 일부 유형에 대한 품질기준 정의 사례는 비정형 콘텐츠에 대한 품질측정 시 측정 기준으로 활용될 수 있는 항목들의 사례가 될 수 있으며, 조직이나 기업·기관의 활용 목적에 따라 보다 다양한 유형의 비정형 콘텐츠에 대해 확대하여 정의하고 사용할 수 있다.

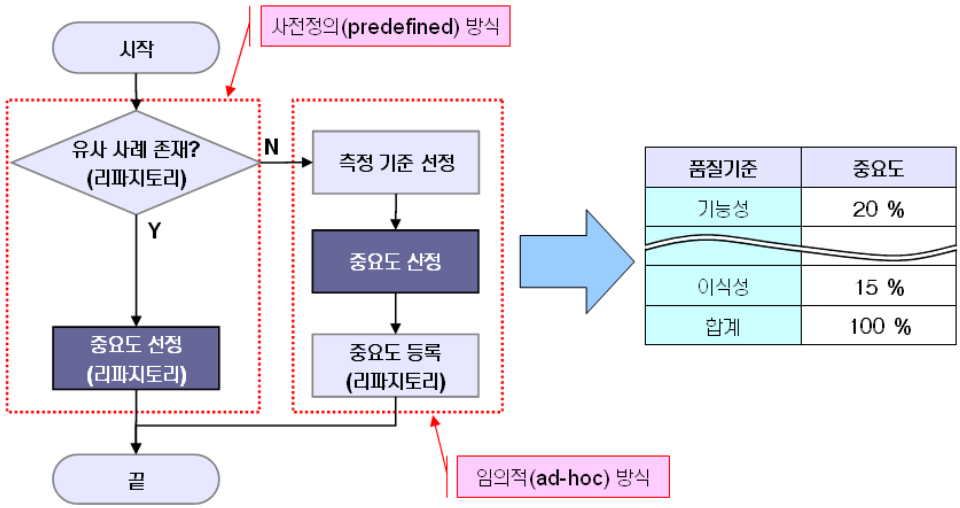
측정 기준의 선정 범위는 품질관리 정책 및 방향에 의해 달라질 수 있으며, 품질측정 목적에 따라서도 품질기준 항목 전체 또는 일부 항목으로 측정 기준의 범위를 선택할 수 있다. 그러나 측정 기준의 선정이 품질측정 시마다 원칙 없이 변경되는 것은 일관성 있는 품질관리에 대해서도 바람직하지 않기 때문에, 품질진단 계획수립 시 장기적인 비전과 목적하에서 명확한 방향이 정의되고 이에 따른 측정기준 선정이 이루어지도록 해야 한다.

측정 기준의 선택은 〈표 2-5〉 ~ 〈표 2-8〉에서 보듯이 비정형 콘텐츠의 유형에 따라 달라질 수 있고, 선정된 측정 기준 또한 진단 대상 콘텐츠의 작성·구축·활용 목적에 따라 측정 기준 간의 중요도에 차이가 있을 수 있다. 예컨대, 특정 지역의 문화를 소개하는 동영상의 경우라면 동영상에 대한 품질기준 항목 중 측정 기준으로서 신뢰성과 사용성이 가장 중요시 될

수 있을 것이다. 이와 같은 측정 기준 간의 가중치 차이를 반영하는 것은 측정 기준에 따른 측정 항목 및 측정 내용들의 측정 결과가 측정 기준의 가중치에 따라 그 신뢰성 및 중요도에서 차이가 있을 수 있어 품질지수 산출 시 이를 반영할 필요가 있기 때문이다.

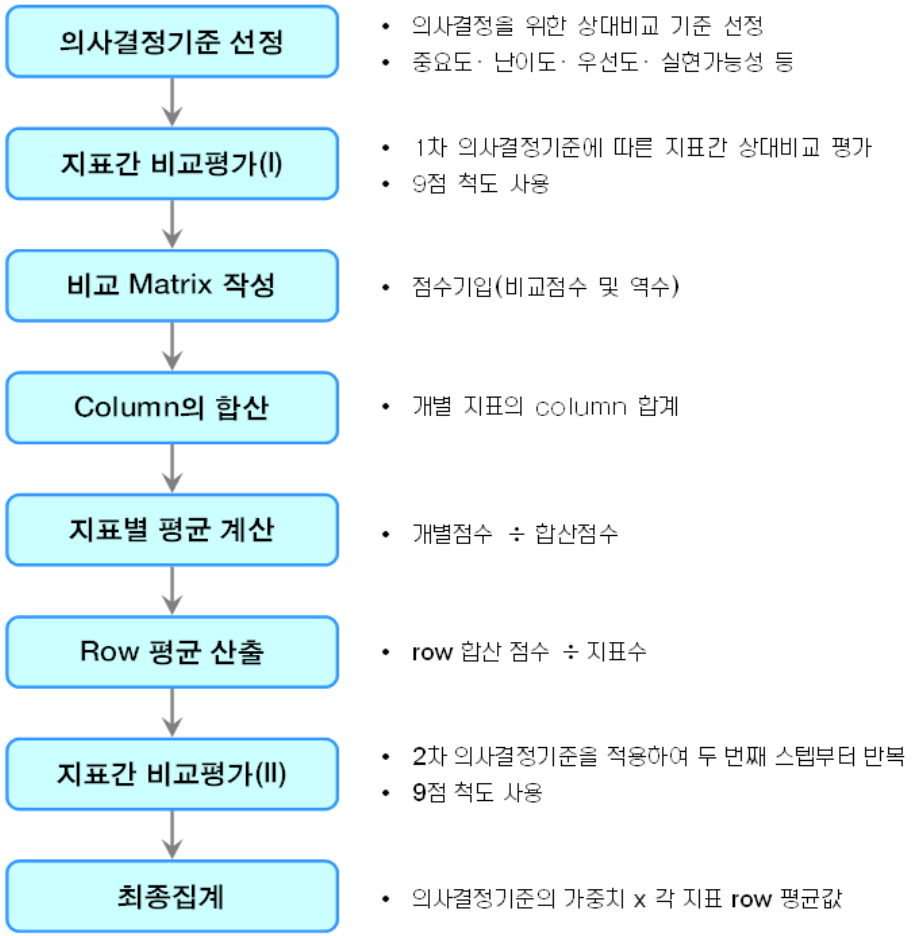
## 2. 중요도 산정

중요도를 산정하여 측정기준 간의 가중치를 정의하는 방법은 사전정의(predefined) 방식이나 임의적(ad-hoc) 방식으로 이루어질 수 있다. 사전정의(predefined) 방식은 진단하려는 대상, 콘텐츠 구축·활용 목적, 진단 목적 등을 고려하여 일치하는 측정 기준으로 작성된 중요도를 선정하여 활용하는 방식으로, 이러한 목적을 위해 리파지토리를 활용하면 매우 효과적일 수 있다. 임의적(ad-hoc) 방식은 사전에 정의된 바가 없어 별도의 중요도를 산정해야 하는 경우 통계적 분석방법을 통해 측정 기준 간의 가중치를 정량적으로 산출하는 방법으로, 품질관리자·품질담당자·콘텐츠담당자 등을 포함한 관련 전문가들의 의견이 적절히 반영되도록 하는 것이 필수적이다. 임의적 방식으로 유용하게 적용될 수 있는 방법의 예로는 AHP(Analytic Hierarchy Process, 계층적 분석법)와 같은 분석적 의사결정방법이 있으며, 때로는 〈표 3-45〉 ~ 〈표 3-48〉에서 제시한 바와 같이 관련 전문가들의 검토에 의해 검증된 판단 기준을 정의하고 이에 따라 중요도를 산정 방식이 있다. 〈그림 4-1〉은 두 가지 방식에 따른 중요도 산정 흐름을 요약한 것이다.




〈그림 4-1〉 중요도 산정방식별 흐름도

임의적(ad-hoc) 방식에 의한 중요도 산정 방법의 하나인 AHP 분석법을 이용한 측정기준 간 가중치 도출 과정의 사례는 〈그림 4-2〉와 같다.




〈그림 4-2〉 AHP 분석법을 이용한 측정기준 간 중요도 산정 절차 사례

## 2.1 의사결정기준 선정

측정기준 간 가중치 산출을 위한 의사결정에 필요한 상대비교 기준을 선정한다. 일반적으로 많이 사용되는 의사결정기준으로는 중요도·난이도·우선도·실현가능성 등이 있다. 동영상 콘텐츠의 품질기준을 측정기준 지표로 하여 가중치 계산을

한다고 가정했을 때 지표 선정은 다음과 같다.

- 지표 1 : 기능성
- 지표 2 : 신뢰성
- 지표 3 : 사용성
- 지표 4 : 효율성
- 지표 5 : 이식성

이와 같은 다섯 가지 지표에 대해 가중치 산출을 위한 의사결정기준으로 중요도와 우선도를 선정하였다고 가정한다.

## 2.2 지표 간 비교평가(I)

선정된 지표 간에 1:1 상대비교를 위한 비교평가척도를 정의한다. 비교평가는 9점 척도 체계를 적용한다.

〈표 4-1〉 AHP 분석을 위한 지표 간 비교평가척도 사례

| 평가척도 | 점수 |
|---|---|
| 지표1이 지표2 보다 극히 중요하다 | 5 |
| 지표1이 지표2 보다 매우 중요하다 | 4 |
| 지표1이 지표2 보다 상당히 중요하다 | 3 |
| 지표1이 지표2 보다 조금 더 중요하다 | 2 |
| 지표1이 지표2 와 동일하게 중요하다 | 1 |
| 지표1이 지표2 보다 조금 덜 중요하다 | 1/2 |
| 지표1이 지표2 보다 상당히 덜 중요하다 | 1/3 |
| 지표1이 지표2 보다 매우 덜 중요하다 | 1/4 |
| 지표1이 지표2 보다 극히 덜 중요하다 | 1/5 |

앞서 선정한 의사결정기준인 중요도와 우선도 중 먼저 중요도에 따른 지표 간 비교평가를 한다고 했을 때, 〈표 4-1〉에 따른 비교평가 결과는 〈표 4-2〉와 같다. 예를 들어, 지표1(기능성)과 지표2(신뢰성)을 비교했을 때 지표1이 지표2 보다 상당히 덜 중요하다면 〈표 4-1〉의 평가척도 점수 기준에 의해 지표1과 지표2의 비교결과 점수는 0.33이 된다는 것이 〈표 4-2〉에 나타나 있다.

〈표 4-2〉 의사결정기준에 따른 지표 간 비교평가 사례

| 의사결정기준 | 지표비교 | | 점수 |
|---|---|---|---|
| | 기준지표 | 비교지표 | |
| 중요도 | 지표( 1 ) | 지표( 2 ) | 0.33 |
| | 지표( 1 ) | 지표( 3 ) | 0.25 |
| | 지표( 1 ) | 지표( 4 ) | 0.5 |
| | 지표( 1 ) | 지표( 5 ) | 2 |
| | 지표( 2 ) | 지표( 3 ) | 0.5 |
| | 지표( 2 ) | 지표( 4 ) | 2 |
| | 지표( 2 ) | 지표( 5 ) | 4 |
| | 지표( 3 ) | 지표( 4 ) | 4 |
| | 지표( 3 ) | 지표( 5 ) | 4 |
| | 지표( 4 ) | 지표( 5 ) | 2 |

## 2.3 비교 Matrix 작성 및 Column 합산

지표 간 비교평가가 완료되면 평가척도에 의하여 비교한 결과를 우상단 측에 기록하고 기록한 결과의 역수를 좌하단 측에 기록하여 〈표 4-3〉과 같이 비교 Matrix를 작성하고 각 Column별 합을 구한다.

〈표 4-3〉 지표 간 비교평가 결과에 따른 비교 Matrix 작성 사례

| 의사결정 기준 | 비교지표 / 기준지표 | 지표1 | 지표2 | 지표3 | 지표4 | 지표5 |
|---|---|---|---|---|---|---|
| 중요도 | 지표1 | 1 | 0.33 | 0.25 | 0.5 | 2 |
| | 지표2 | 3 | 1 | 0.5 | 2 | 4 |
| | 지표3 | 4 | 2 | 1 | 4 | 4 |
| | 지표4 | 2 | 0.5 | 0.25 | 1 | 2 |
| | 지표5 | 0.5 | 0.25 | 0.25 | 0.5 | 1 |
| Column 합계 | | 10.5 | 4.083 | 2.25 | 8 | 13 |

## 2.4 지표별 평균 계산 및 Row 평균 산출

지표 간 비교평가를 통해 비교 Matrix가 작성되면 각 지표의 값을 컬럼별 지표 합산값으로 나누어 지표별 평균을 구하고, 각 Row에 대해 지표별 평균의 합을 컬럼수로 나눈 Row 평균을 구한다.

〈표 4-4〉 비교 Matrix 계산 사례

| 의사결정 기준 | 비교지표 / 기준지표 | 지표1 | 지표2 | 지표3 | 지표4 | 지표5 | 각 Row 평균 |
|---|---|---|---|---|---|---|---|
| 중요도 | 지표1 | 0.095238095 | 0.081632653 | 0.111111111 | 0.0625 | 0.153846154 | 0.100865603 |
| | 지표2 | 0.285714286 | 0.244897959 | 0.222222222 | 0.25 | 0.307692308 | 0.262105355 |
| | 지표3 | 0.380952381 | 0.489795918 | 0.444444444 | 0.5 | 0.307692308 | 0.42457701 |
| | 지표4 | 0.19047619 | 0.12244898 | 0.111111111 | 0.125 | 0.153846154 | 0.140576487 |
| | 지표5 | 0.047619048 | 0.06122449 | 0.111111111 | 0.0625 | 0.076923077 | 0.071875545 |

## 2.5 지표간 비교평가(II)

하나의 의사결정기준에 대해 지표 간 비교평가를 통해 비교 Matrix를 작성하고 계산하였던 방법을 또 다른 의사결정기준

에 대해 적용하여 〈표 4-2〉 ~ 〈표 4-4〉의 과정에 이르는 계산을 반복한다.

## 2.6 최종 집계 및 가중치 산정

두 개의 의사결정기준의 상대적 가중치를 결정한 다음 각각의 가중치에다 Row별 평균값을 곱하여 최종가중치를 산정한다. 위 사례에서 중요도와 우선도의 가중치 비율을 0.4 : 0.6으로 가정하면 최종 가중치 산정 결과는 〈표 4-5〉에서 보는 바와 같다.

〈표 4-5〉 최종 가중치 산정 결과 사례

| | 중요도에 따른 ROW 평균 (①) | 우선도에 따른 ROW 평균 (②) | 최종 가중치 (①×0.4)+(②×0.6) | |
|---|---|---|---|---|
| | CI=0.0286 | CI=0.0392 | | |
| 지표1 (기능성) | 0.101 | 0.070 | 0.083 | ≈ 0.1 |
| 지표2 (신뢰성) | 0.262 | 0.294 | 0.281 | ≈ 0.3 |
| 지표3 (사용성) | 0.425 | 0.413 | 0.418 | ≈ 0.4 |
| 지표4 (효율성) | 0.141 | 0.127 | 0.133 | ≈ 0.1 |
| 지표5 (이식성) | 0.072 | 0.095 | 0.085 | ≈ 0.1 |

AHP 분석법에서 중요한 것은 지표 간의 비교평가에 있어서 논리적 일관성(logical consistency)를 유지하는 것이다. 이를 위해 지표 간 비교평가에 관련 전문가의 의견이 반영되어야 하며, 한 명 보다는 다수의 전문가들이 평가한 결과의 평균이 최종 가중치로 선정되는 것이 바람직하다. 지표 간 비교평가에 있어서 응답의 신뢰성 검증 및 논리적 일관성 유지에 대한 평가는 일관성 지수(Consistency Index; CI)를 구하여 확인할 수 있다. 응답자가 논리적으로 모순을 유발하게 되면 (예: A는 B보다 중요하고, B는 C보다 중요하다고 말해 놓고,

A는 C보다는 덜 중요하다 라고 하는 경우) CI가 증가하는데, 통상 CI 값이 0.1 이상이 되면 응답자의 답변을 신뢰할 수 없다고 본다. 이 때문에 일관성 지수(CI)는 응답에 대한 논리적 모순을 검출하기 위해 꼭 산정해 보아야 한다. CI를 계산하는 방법은 다음과 같다.

일관성지수 $CI = \frac{\lambda_{max} - n}{n - 1}$

여기서 $\lambda_{max}$ 는 최대고유치(Principal Eigen value)이고, n은 행렬의 차원(지표수)으로, n x n 정방행렬 [A]와 n x 1 가중치행렬 [W]를 곱하면 새로운 n x 1 가중벡터행렬 [Y]가 산정되는데, 이 가중벡터행렬의 구성요소 $Y_1$ … $Y_n$ 과 가중치 $W_1$ … $W_n$ 을 이용하여 다음과 같이 $\lambda$max 를 계산한다.

[A] X [W] = [Y] 일 때,

$\lambda_{max} = (Y_1/W_1 + Y_2/W_2 + \cdots + Y_n/W_n) / n$

이와 같은 CI 계산 방법에 의거하여 위 사례에서 중요도와 우선도에 따른 가중치 계산에 대한 CI를 산정해 보면 〈표 4-5〉에 나타낸 것처럼 모두 0.1 보다 작기 때문에 지표 간 비교평가에 대한 논리적 일관성이 유지되고 있고, 이에 따라 최종 가중치 산정 결과가 신뢰할 수 있음을 확인할 수 있다. AHP 분석법을 적용하는데 있어서 한 가지 주의할 점은 비교항목이 너무 많으면 복잡성이 크게 증가하여 적용이 쉽지 않으며, 15개가 넘으면 가중치 산출이 불가능하기 때문에 비교항목을 선정하는데 있어서 이 점을 반드시 고려해야 한다. 이 때문에 AHP 분석법에 의한 측정기준 간 가중치 산정 시는 앞서 예로 든 비정형 콘텐츠의 품질기준 중 주특성을 비교항목으로 선택하여 가중치를 산출하는 것이 적당하다.

〈표 4-6〉과 〈표 4-7〉은 AHP 분석법에 의한 중요도 산정으로 측정기준 간 가중치를 도출한 사례이다.

〈표 4-6〉 AHP 분석법에 따른 콘텐츠 유형별 최종 가중치 산정 결과 사례(1)

| 품질기준 / 진단대상 | 기능성 | 신뢰성 | 사용성 | 효율성 | 이식성 | 총합 |
|---|---|---|---|---|---|---|
| 동영상 | 0.2 | 0.3 | 0.3 | 0.1 | 0.1 | 1 |
| 이미지 | 0.1 | 0.2 | 0.3 | 0.3 | 0.1 | 1 |
| 사운드 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 1 |

〈표 4-7〉 AHP 분석법에 따른 콘텐츠 유형별 최종 가중치 산정 결과 사례(2)

| 품질기준 / 진단대상 | 해상도 | 정확성 | 완전성 | 일관성 | 총합 |
|---|---|---|---|---|---|
| GIS | 0.2 | 0.3 | 0.1 | 0.4 | 1 |

## 3. 측정 항목의 작성

비정형 콘텐츠에 대한 품질측정에 있어서 측정 항목은 측정기준에 따른 주요 품질 항목을 지칭하는 것으로, 콘텐츠 유형과 품질기준에 따라 달라질 수 있다. 〈표 4-8〉은 동영상 콘텐츠에 대한 측정 항목 구성 사례이다.

〈표 4-8〉 동영상 콘텐츠에 대한 측정항목 구성 사례

| 콘텐츠 유형 | 품질기준 | 세부기준 | 측정 항목 |
|---|---|---|---|
| 동영상 | 기능성 | 정확성 | 부가요소 정확성(자막 정확성) |
| | | | 부가요소 정확성(사운드 정확성) |
| | | | 메타데이터 연결 정확성 |
| | | 적절성 | 목적에 대한 내용 부합성 |
| | | | 운용적절성(비디오압축코덱) |
| | | | 운용적절성(오디오압축코덱) |
| | | | 운용적절성(초당프레임수) |
| | | | 운용적절성(사운드채널) |
| | | | 운용적절성(화면비율) |
| | | | 운용적절성(Running Time) |
| | | 상호운용성 | 사운드 동기화 |
| | | | 자막 동기화 |
| | | 기능순응성 | 기능성에 대한 규격화 여부 |
| | | | 규격에 대한 공유와 숙지 |
| | | | 규격에 대한 준수 정도 |
| | 신뢰성 | 성숙성 | 기준환경에서의 결함발생 정도 |
| | | | 결함 발생에 대한 대응성 |
| | | 신뢰순응성 | 신뢰성 관련 기준 환경 및 적용기준 규격화 (모니터, 컴퓨터 최소사양, 실행환경설정표준 등) |
| | | | 신뢰성 관련 규격의 공유와 숙지 |
| | | | 규격에 대한 준수 정도 |
| | 사용성 | 이해성 | 영상 끊김 |
| | | | 영상인식 만족도(영상 선명도) |
| | | | 영상인식 만족도(자막 선명도) |
| | | | 음향인식 만족도 |
| | | 친밀성 | 포맷 친숙성 |
| | | | 기준 환경의 친숙성 |
| | | | 기준 환경의 적절성 |
| | | 사용순응성 | 사용성에 대한 규격화 여부 |
| | | | 규격에 대한 공유와 숙지 |
| | | | 규격에 대한 준수 정도 |
| | 효율성 | 시간효율성 | 응답속도 |
| | | | 버퍼링 |
| | | 효율순응성 | 효율성에 대한 규격화 여부 |
| | | | 규격에 대한 공유와 숙지 |
| | | | 규격에 대한 준수 정도 |
| | 이식성 | 적응성 | 운영환경 및 플레이어 호환성 |
| | | 공존성 | Running시 타SW 영향 여부 |
| | | 이식순응성 | 이식성에 대한 규격화 여부 |
| | | | 규격에 대한 공유와 숙지 |
| | | | 규격에 대한 준수 정도 |

측정 항목은 선정된 측정 기준에 대해서 실제 품질측정을 수행할 구체적인 항목을 정의한 것으로, 〈표 4-8〉의 측정항목 구성 사례에 보는 것처럼 선정된 측정 기준에 대해 각 조직이나 기업・기관에서 필요한 항목들로 재구성하여 사용할 수 있으며, 측정 기준의 가중치에 따라 측정 항목의 최종 가중치 또한 달라질 수 있다. 즉, 측정 항목이 선정되면 개별 측정 항목에 대한 중요도를 평가하여 해당 측정 기준에 대한 가중치를 추가 반영함으로써 하나의 측정 기준에 대한 측정 항목들 중에서 다시 임의의 측정 항목이 갖는 최종 가중치를 결정하게 된다. 이와 같은 가중치의 중첩 반영을 통해서 측정 항목에 콘텐츠 특성과 품질측정 목적이 반영되게 된다. 측정 항목의 중요도 평가에 따른 최종 가중치를 결정하는데 있어서 고려되는 요소는 다음과 같이 정의될 수 있다.

〈표 4-9〉 측정 항목의 최종 가중치 결정에 대한 고려 요소

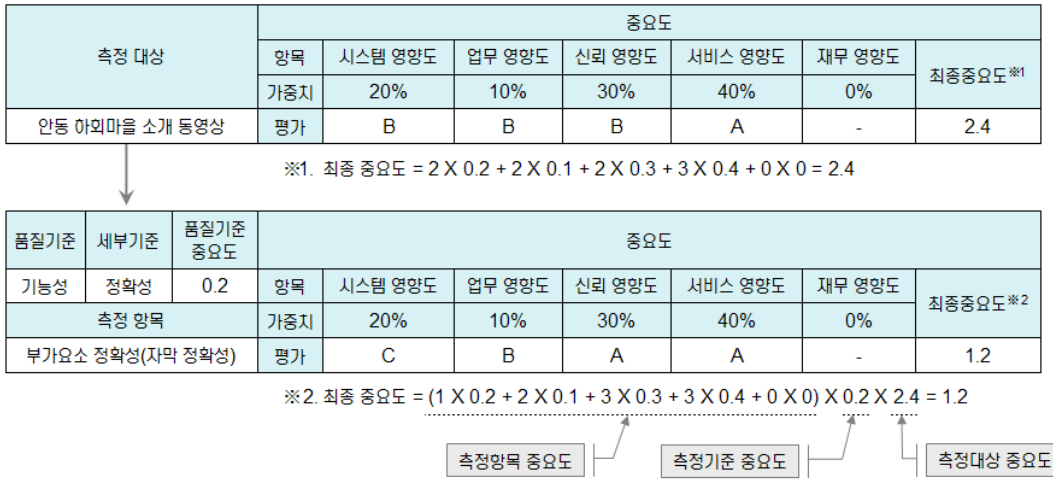
| 고려 요소 | 설 명 |
|---|---|
| 측정 기준의 중요도 | 앞에서 설명한 측정 기준의 중요도 산정에 의해 도출된 측정기준 간의 가중치이다. |
| 측정 대상의 중요도 | 품질측정을 하고자 하는 진단 대상 콘텐츠의 중요도로서, 측정 항목의 중요도 평가와 동일한 기준 및 방법에 의해 중요도 값을 산출한다. |
| 측정 항목의 중요도 | 측정 기준에 대한 구체적인 품질 항목들로, 개별 측정 항목에 대해 시스템적 요소와 업무적 요소의 영향도 평가를 통해 해당 측정 항목의 중요도를 산정한다. |

측정 항목의 중요도 평가는 비교할 항목이 많기 때문에 AHP 분석법과 같은 분석적 의사결정방법을 적용하는 것보다 품질 문제 발생 시의 영향범위 검토를 통해 도출한 평가 요소를 적용하여 중요도를 평가하는 것이 더 효율적이다. 측정 항목에 대한 중요도 평가 요소는 〈표 4-10〉에서 보는 바와 같다.

〈표 4-10〉 비정형 콘텐츠의 측정 항목에 대한 중요도 평가 요소

| 비정형 콘텐츠에 대한 품질문제 유형 | 영향 범위 | 평가 요소 |
|---|---|---|
| 콘텐츠 실행이 안됨 | 품질문제 발생 시 시스템에 영향을 미침 | 시스템영향도 |
| 콘텐츠 실행은 되나 인식이 어려움 | 품질문제 발생 시 업무 수행에 영향을 미침 | 업무영향도 |
| 원하는 콘텐츠가 아님 | 품질문제 발생 시 서비스 제공자의 신뢰성에 영향을 미침(예: 대민 서비스 신뢰도) | 신뢰영향도 |
| 원하는 콘텐츠를 찾기가 어려움 | 품질문제 발생 시 고객 서비스에 영향을 미침 (예: 사적지DB - 우리고장 사적지 소개) | 서비스영향도 |
| 원하는 콘텐츠가 제때에 서비스 안됨 | 품질문제 발생 시 재무적 영향이 발생함 (예: 유료 콘텐츠 서비스) | 재무영향도 |
| 너무 느림 | | |
| 실행시 시스템이 느려짐 | | |
| 유료서비스인데 제대로 실행이 안됨 | | |
| 유료서비스인데 콘텐츠 수준이 낮음 | | |
| 장애복구 후 콘텐츠 서비스가 안됨 | | |

〈표 4-10〉에서 도출된 평가 요소에 대해 정형 텍스트 데이터에 대한 업무규칙 중요도 평가에서와 같이 각 평가 요소별로 그 영향 정도에 따라 상(A)・중(B)・하(C)로 구분하고, 각각 3・2・1 점을 배점하며, 각 평가 요소별로 적용 가중치를 고려하여 평가된 배점의 곱을 합산한 값으로 측정 항목의 중요도를 산정하고, 여기에 동일한 방법으로 산출한 측정 대상의 중요도와 측정 기준의 가중치를 곱하여 측정 항목의 최종 가중치를 산출한다. 〈그림 4-3〉과 〈표 4-11〉은 이와 같은 방법에 따라 측정 항목의 최종 가중치를 평가하는 사례이다.




〈그림 4-3〉 측정 항목의 최종 가중치 평가 사례

〈표 4-11〉 다중 측정 항목에 대한 최종 가중치 평가 사례

| 콘텐츠 유형 | 품질 기준 | 품질 기준 중요도 | 세부 기준 | 측정 항목 | 중요도 | | | | | | | | | | 최종 중요도 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 시스템 영향도 | | 업무 영향도 | | 신뢰 영향도 | | 서비스 영향도 | | 재무 영향도 | | |
| | | | | | 가중치 | 평가 | 가중치 | 평가 | 가중치 | 평가 | 가중치 | 평가 | 가중치 | 평가 | |
| 동영상 | 기능성 | 0.4 | 정확성 | 부가요소 정확성(자막 정확성) | 20% | C | 10% | B | 30% | A | 40% | A | 0% | - | 2.5 |
| | | | | 부가요소 정확성(사운드 정확성) | 20% | B | 10% | C | 30% | C | 40% | B | 0% | - | 1.6 |
| | | | 적절성 | 운용적절성(비디오압축 코덱) | 30% | A | 10% | B | 20% | C | 40% | A | 0% | - | 2.5 |
| | | | 기능순응성 | 규격화 | 20% | A | 20% | B | 30% | C | 30% | B | 0% | - | 1.9 |
| | 효율성 | 0.6 | 시간효율성 | 응답속도 | 40% | A | 10% | A | 20% | B | 30% | A | 0% | - | 2.8 |
| | | | 효율순응성 | 규격화 | 20% | A | 20% | B | 30% | B | 30% | A | 0% | - | 2.5 |
| | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |

## 4. 측정 내용의 작성

측정 항목이 선정되면 각 측정 항목에 대해 실제로 품질측정을 수행할 내용을 작성하여 체크리스트를 구성하게 된다. 측정 내용은 시나리오 구성 형태로 작성하는 것이 효율적인데, 이는 측정 수행 내용을 기술하고, 이에 대한 측정 결과를 5점 척도에 따라 선택할 수 있도록 예상되는 결과를 제시하며, 각각의 측정 내용에 대한 결과 판단 기준과 측정 방법으로 구성하여, 측정 수행자가 측정 내용별 측정 방법에 따라 측정을 수행한 후 결과 판단 기준에 의거하여 측정된 결과를 측정 내용별 5점 척도 구성 보기 중에서 선택함으로써 측정이 이루어지게 되는 구성이다. 이와 같은 측정 내용은 측정 항목별로 하나 이상 도출될 수 있다. 〈표 4-12〉는 이와 같은 방법에 따라 작성된 비정형 콘텐츠에 대한 품질측정 체크리스트의 사례이다.

〈표 4-12〉 비정형 콘텐츠의 품질측정 체크리스트 구성 사례

| 측정대상 | 측정 대상 중요도 | 품질기준 | 품질기준 중요도 | 세부기준 | 측정항목 | 측정항목 중요도 | 측정내용 | 측정기준 | 측정방법 |
|---|---|---|---|---|---|---|---|---|---|
| △△사적지 소개동영상 | 2.4 | 기능성 | 0.4 | 정확성 | 부가요소 정확성(자막 정확성) | 2.5 | **자막은 맞춤법 표기에 따라 작성 되었는가?**<br>5 : 100% 맞춤법 표기 준수<br>4 : 98% 이상 맞춤법 표기 준수<br>3 : 95% 이상 맞춤법 표기 준수<br>2 : 90% 이상 맞춤법 표기 준수<br>1 : 90% 미만 맞춤법 표기 준수 | 100% (콘텐츠 작성표준 지침) | Sampling |
| | | | | | 부가요소 정확성(사운드 정확성) | 1.6 | **나레이션 시나리오와 사운드 내용은 일치하는가?**<br>5 : 100% 일치<br>4 : 98% 이상 일치<br>3 : 95% 이상 일치<br>2 : 90% 이상 일치<br>1 : 90% 미만 일치 | 100% (콘텐츠 작성표준 지침) | Sampling 누락, 잘못읽음, 발음부정확 건수 |
| | | | | 적절성 | 운용적절성 (비디오압축코덱) | 2.5 | **비디오 압축 코덱은 표준을 준수하고 있는가?**<br>5 : 모두 준수<br>4 : 98% 이상 준수<br>3 : 95% 이상 준수<br>2 : 90% 이상 준수<br>1 : 90% 미만 준수 | 100% (콘텐츠 작성표준 지침) | 전수 검사 |
| | | | | 기능순응성 | 규격화 | 1.9 | **기능성 관련 항목에 대한 표준 지침이 있는가?**<br>5 : 모든 항목에 대해 표준이 문서화되어 있음<br>4 : 일부 항목에 대해 표준이 문서화되어 있음<br>3 : 모든 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>2 : 일부 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>1 : 표준이 없음 | 기능성 관련 항목의 표준이 모두 문서로 명시되어 있어야 함 | 기능성 관련 항목의 표준이 문서로 명시되어 있는지 확인 |
| | | 효율성 | 0.6 | 시간효율성 | 응답속도 | 2.8 | **선택한 동영상이 기준 시간 내에 로딩되는가?**<br>5 : 모두 만족함<br>4 : 90% 이상 만족함<br>3 : 80% 이상 만족함<br>2 : 70% 이상 만족함<br>1 : 70% 미만 만족함 | 〈 1초 | 전수검사 |
| | | | | 효율순응성 | 규격화 | 2.5 | **효율성 관련 항목에 대한 표준 지침이 있는가?**<br>5 : 모든 항목에 대해 표준이 문서화되어 있음<br>4 : 일부 항목에 대해 표준이 문서화되어 있음<br>3 : 모든 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>2 : 일부 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>1 : 표준이 없음 | 효율성 관련 항목의 표준이 모두 문서로 명시되어 있어야 함 | 효율성 관련 항목의 표준이 문서로 명시되어 있는지 확인 |

# 제3절 품질측정 및 품질 지수의 산출

비정형 콘텐츠에 대한 품질측정 체크리스트를 작성하여 품질 측정을 수행하게 되면 측정 내용 구성에 따른 측정 결과가 얻어지게 된다. 정형 텍스트에 대한 품질측정은 업무규칙이 측정 내용이 되어 각 업무규칙 적용에 따른 오류율이 도출되고, 이 오류율에 대해 가중치 적용을 통해 최종적인 품질지수를 도출하는 방식으로 데이터의 품질현황을 평가하게 되지만, 비정형 콘텐츠에 대한 품질측정의 경우는 측정 항목에 대한 실제 측정 내용으로 구성된 체크리스트를 작성하고 이를 이용하여 품질측정을 하게 되는데, 이때 측정 결과가 5점 척도로 구성된 보기 중에서 선택되므로 각 측정 내용들은 모두 1 ~ 5점 사이의 값을 갖게 된다. 그러므로 이들 각 측정 내용에 대해 평가된 점수를 취합하여 환산하게 되면 측정 대상 콘텐츠에 대한 품질점수를 얻게 되며, 이 품질점수에 대해 2차적인 가공을 통해 콘텐츠 유형별 품질지수로 전환하는 방법을 사용한다. 콘텐츠 유형별로 평가된 품질지수는 동영상・사운드・이미지 등과 같은 여러 유형의 비정형 콘텐츠에 대한 품질측정이 수행된 경우 유형별 품질지수의 평균이나 또는 콘텐츠 유형별 가중치를 적용하여 각 유형별 품질지수를 총품질지수라는 하나의 수치로 표현할 수 있다.

비정형 콘텐츠에 대한 품질지수와 총품질지수 산출은 비정형 콘텐츠에 대한 품질현황을 정형 텍스트 데이터의 경우에서와 마찬가지로 수치적으로 가시화(Visualization)하는 효과를 얻을 수 있기 때문에, 이를 통해 현행의 품질 수준과 품질 개선 정도를 표현하는데 용이하고, 품질관리를 효율화할 수 있는 객관적인 근거를 제시한다는 측면에서 매우 큰 의미를 갖는다.

## 1. 품질점수의 산출

앞서 소개한 체크리스트를 이용한 비정형 콘텐츠의 품질측정 결과를 용이하게 평가하기 위해서는 체크리스트 구성 항목에 측정결과를 기록하기 위한 항목이 추가되어야 한다. 〈표 4-13〉의 사례에서 '현수준'이라 표기한 항목은 이와 같은 목적으로 추가된 측정결과를 기재하는 항목이다. 앞서 언급한 것처럼 측정 결과는 1~5점 사이의 값으로 기록되기 때문에 해당 측정 기준에 대한 측정 항목 및 측정 내용들 전체에 대한 측정 결과점수를 취합하여 해당 측정 기준에 대한 취득 점수를 100점 기준으로 표현할 수 있는데, 이를 '품질점수'라고 한다. 품질점수는 다음과 같이 계산한다.

품질점수 $Q_S = \frac{\sum R_i \times 20}{C_q}$

$Q_S$ : 측정기준별 품질점수

$\Sigma R_i$ : 측정기준별 측정결과점수의 합

$C_q$ : 측정기준별 측정문항수

〈표 4-13〉은 앞서 제시한 〈표 4-12〉의 체크리스트 사례에 현수준 항목을 추가하여 측정결과를 기록한 사례이며, 사례에 제시된 측정결과점수를 토대로 하여 앞에서 제시한 수식을 통해 품질점수를 계산하는 사례는 다음과 같다.

〈표 4-13〉 체크리스트에 대한 측정결과 기록 사례

| 측정대상 | 측정 대상 중요도 | 품질 기준 | 품질 기준 중요도 | 세부기준 | 측정항목 | 측정항목 중요도 | 측정내용 | 측정기준 | 측정방법 | 현수준 |
|---|---|---|---|---|---|---|---|---|---|---|
| △△사적지 소개동영상 | 2.4 | 기능성 | 0.4 | 정확성 | 부가요소 정확성 (자막 정확성) | 2.5 | **자막은 맞춤법 표기에 따라 작성 되었는가?**<br>5 : 100% 맞춤법 표기 준수<br>4 : 98% 이상 맞춤법 표기 준수<br>3 : 95% 이상 맞춤법 표기 준수<br>2 : 90% 이상 맞춤법 표기 준수<br>1 : 90% 미만 맞춤법 표기 준수 | 100% (콘텐츠 작성표준 지침) | Sampling | 4 |
| | | | | | 부가요소 정확성 (사운드 정확성) | 1.6 | **나레이션 시나리오와 사운드 내용은 일치하는가?**<br>5 : 100% 일치<br>4 : 98% 이상 일치<br>3 : 95% 이상 일치<br>2 : 90% 이상 일치<br>1 : 90% 미만 일치 | 100% (콘텐츠 작성표준 지침) | Sampling 누락, 잘못읽음, 발음부정확 건수 | 3 |
| | | | | 적절성 | 운용 적절성 (비디오 압축코덱) | 2.5 | **비디오 압축 코덱은 표준을 준수하고 있는가?**<br>5 : 모두 준수<br>4 : 98% 이상 준수<br>3 : 95% 이상 준수<br>2 : 90% 이상 준수<br>1 : 90% 미만 준수 | 100% (콘텐츠 작성표준 지침) | 전수 검사 | 5 |
| | | | | 기능 순응성 | 규격화 | 1.9 | **기능성 관련 항목에 대한 표준 지침이 있는가?**<br>5 : 모든 항목에 대해 표준이 문서화되어 있음<br>4 : 일부 항목에 대해 표준이 문서화되어 있음<br>3 : 모든 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>2 : 일부 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>1 : 표준이 없음 | 기능성 관련 항목의 표준이 모두 문서로 명시되어 있어야 함 | 기능성 관련 항목의 표준이 문서로 명시되어 있는지 확인 | 4 |
| | | 효율성 | 0.6 | 시간 효율성 | 응답속도 | 2.8 | **선택한 동영상이 기준 시간 내에 로딩되는가?**<br>5 : 모두 만족함<br>4 : 90% 이상 만족함<br>3 : 80% 이상 만족함<br>2 : 70% 이상 만족함<br>1 : 70% 미만 만족함 | 〈 1초 | 전수검사 | 3 |
| | | | | 효율 순응성 | 규격화 | 2.5 | **효율성 관련 항목에 대한 표준 지침이 있는가?**<br>5 : 모든 항목에 대해 표준이 문서화되어 있음<br>4 : 일부 항목에 대해 표준이 문서화되어 있음<br>3 : 모든 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>2 : 일부 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>1 : 표준이 없음 | 효율성 관련 항목의 표준이 모두 문서로 명시되어 있어야 함 | 효율성 관련 항목의 표준이 문서로 명시되어 있는지 확인 | 4 |

측정기준 '기능성'에 대한 품질점수 = (4+3+5+4) X 20 / 4 = 80

측정기준 '효율성'에 대한 품질점수 = (3+4) X 20 / 2 = 70

## 2. 품질지수의 산출

품질점수로부터 품질지수를 산출하는 방법은 해당 측정기준에 대한 품질점수에 앞서 제시한 측정기준별 가중치를 곱하여 측정기준의 중요도에 따른 해당 품질점수의 기여도에 해당하는 가중품질점수를 구하고, 이들을 합산함으로써 해당 콘텐츠 유형에 대한 품질지수를 계산한다. 품질지수의 계산 방법을 요약하면 다음과 같다.

품질지수 $QI = \sum(Q_S \times Wt_b)$

$QI$ : 해당 콘텐츠 유형에 대한 품질지수

$Q_S$ : 해당 콘텐츠 유형의 측정기준별 품질점수

$Wt_b$ : 해당 콘텐츠 유형의 측정기준별 가중치

〈표 4-14〉는 이 수식에 따른 품질지수 계산 사례를 보여준다.

〈표 4-14〉 비정형 콘텐츠의 품질지수 산출 사례

| 대상 | 측정기준 | 가중치 | 품질점수 | 가중품질점수 | 품질지수 |
|---|---|---|---|---|---|
| 동영상 | 기능성 | 0.2 | 98.5 | 19.7 | 98.4 |
| | 신뢰성 | 0.3 | 99.0 | 29.7 | |
| | 사용성 | 0.3 | 98.7 | 29.61 | |
| | 효율성 | 0.1 | 95.2 | 9.52 | |
| | 이식성 | 0.1 | 98.9 | 9.89 | |

〈표 4-14〉에서 보면 각 측정기준별 가중치와 품질점수를 곱하여 가중품질점수를 산출하고, 다시 이들 가중품질점수를 합산하여 해당 콘텐츠 유형에 대한 품질지수를 계산하였다.

## 3. 총품질지수의 산출

콘텐츠 유형별 품질지수 산출 결과를 토대로 콘텐츠 전체에 대한 품질 수준을 하나의 수치로 표현하기 위한 방법이 총품질지수이다. 총품질지수는 각 조직이나 기업・기관의 품질관

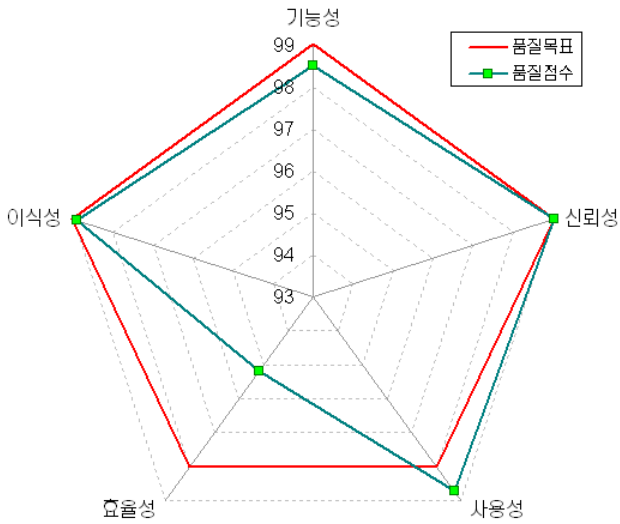
리 수준 및 목적에 따라 비정형 콘텐츠 전체의 품질 수준을 표현하는 수단으로 사용될 수도 있고, 또는 정형 텍스트 데이터에 대한 품질지수와 비정형 콘텐츠에 대한 품질지수 산출 결과를 모두 망라한 종합품질지수의 개념으로도 활용될 수 있다.

품질관리 목적에 따라 품질관리 대상이 되는 모든 유형의 콘텐츠(텍스트 포함)에 대해 다시 가중치를 적용하여 총품질지수를 산출할 수도 있고, 콘텐츠 유형별 품질지수의 단순 평균으로 총품질지수를 사용할 수도 있다. 중요한 것은 총품질지수를 어떻게 산출하였는가 하는 것보다 총품질지수라는 단일값으로 표현된 수치를 이용하여 얼마나 효율적으로 품질수준을 관리하고 향상시키는데 활용할 수 있는가 하는 점이다. 〈표 4-15〉는 총품질지수를 산출하는 사례를 보여준다.

〈표 4-15〉 총품질지수 산출 사례

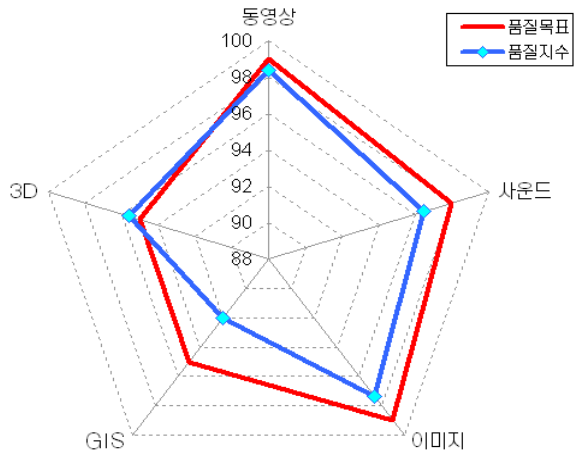
| 대상 | 측정기준 | 가중치 | 품질점수 | 가중품질점수 | 품질지수 | 총품질지수 |
|---|---|---|---|---|---|---|
| 동영상 | 기능성 | 0.2 | 98.5 | 19.7 | 98.4 | 96.0 |
| | 신뢰성 | 0.3 | 99.0 | 29.7 | | |
| | 사용성 | 0.3 | 98.7 | 29.61 | | |
| | 효율성 | 0.1 | 95.2 | 9.52 | | |
| | 이식성 | 0.1 | 98.9 | 9.89 | | |
| 사운드 | … | … | … | … | 96.5 | |
| 이미지 | … | … | … | … | 97.4 | |
| GIS | … | … | … | … | 92.0 | |
| 3D | … | … | … | … | 95.6 | |

품질지수와 총품질지수를 이용하여 현행 품질 수준에 대한 측정 결과를 가시화하면 목표 대비 현행 품질수준의 차이를 용이하게 비교하여 표현할 수 있으며, 집중적인 관리 및 개선이 필요한 부분과 전체적인 품질수준의 변화가 긍정적인 방향으로 가고 있는지 여부 등을 용이하게 파악할 수 있다. 〈그림 4-4〉는 〈표 4-15〉의 사례로부터 측정기준별 품질점수에 대한 현행 수준을 표현한 것이다.



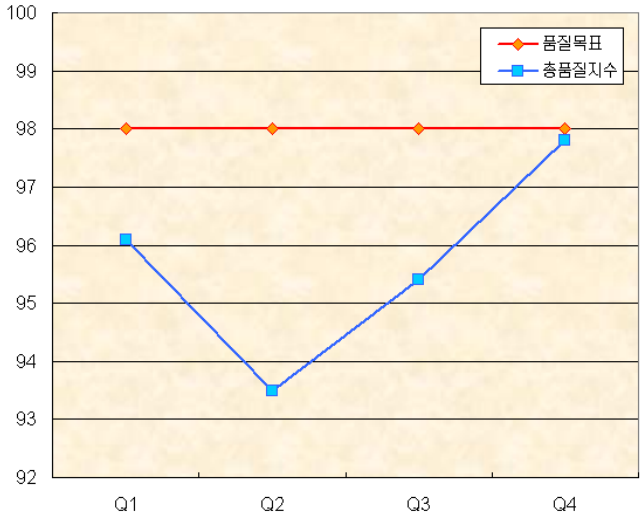

〈그림 4-4〉 측정기준별 목표 대비 품질점수 비교 사례

또한 각 콘텐츠 유형별 품질측정에 의해 도출된 품질지수를 콘텐츠 유형 간에 비교할 수 있는데, 이 경우 역시 콘텐츠 유형별로 목표 대비 현행 수준의 비교를 함으로써 집중 관리 및 개선이 필요한 콘텐츠 유형을 용이하게 설명할 수 있게 된다. 〈그림 4-5〉는 이와 같은 콘텐츠 유형 간 품질지수 비교에 대한 사례를 보여준다.




〈그림 4-5〉 콘텐츠 유형간 목표 대비 품질지수 비교 사례

다양한 콘텐츠 유형에 대한 품질지수를 하나의 수치로 단일화하여 표현한 총품질지수는 품질수준의 변화를 추적하는데 용이한 수단을 제공한다. 품질측정 시마다 산출한 총품질지수의 변화추이를 도식화해 보면 품질수준의 변화를 용이하게 파악할 수 있기 때문에 현재의 품질관리 방법이 효율적인지 개선할 필요가 있는지에 대한 판단까지 가능하다. 〈그림 4-6〉은 총품질지수에 대한 변화추이를 표현한 사례이다.




〈그림 4-6〉 총품질지수에 대한 변화추이 사례

## 4. 오류율 측정

콘텐츠 유형별, 품질기준별 품질측정 내용 중 오류율 측정이 가능한 부분에 대해서는 정형 텍스트 데이터에 대한 오류율 측정과 마찬가지로 오류율을 적용할 수 있다. 그러나 정형 텍스트 데이터에 대한 오류율은 다량의 데이터 중에서 일부에 발생한 오류에 대한 비율이기 때문에 그 수치가 적게 나타나지만, 비정형 콘텐츠에 대해 오류율을 적용할 경우 모집단의 크기 자체가 정형 텍스트 데이터와 비교할 수 없을 정도로 작은 경우가 대부분이기 때문에 정형 텍스트에 대해서와 동일한 개념으로 오류율을 사용할 경우 오해나 왜곡의 가능성이 커질 우려가 있다. 그러므로 비정형 콘텐츠에 대한 오류율은 참고사항의 성격이 강하다고 보아야 할 것이다.

비정형 콘텐츠에 대한 오류율 측정은 체크리스트에 측정한 총건수, 오류건수, 오류율, 가중오류율 항목을 추가하여 측정할 수 있다. 총건수는 해당 측정 내용을 조사한 전체 건수이

고, 오류건수는 총건수 중에서 오류가 발생한 건수로, 여기서 오류는 측정 내용에 기재된 보기 문항 중 기대치에 어긋나는 경우로 볼 수 있다. 오류율은 총건수에 대한 오류건수의 비율이고, 가중오류율은 오류율에 측정항목 중요도를 곱한 결과 값으로, 측정항목의 중요도에 비추어 해당 측정 내용의 오류율이 그만큼 더 또는 덜 중대한 크기로 부각되도록 하기 위함이다. 가중오류율이 산출되면 이들의 합을 측정기준이나 진단 대상 콘텐츠에 대한 측정 내용 문항수로 나누어 가중평균오류율을 얻을 수 있다. 이를 정리하면 다음과 같다.

측정내용별 오류율 $e_i = \frac{c_e}{c_t}$

( $c_e$ : 오류건수, $c_t$ : 총건수 )

측정내용별 가중오류율 $e_{wi} = e_i \times Wt_{qi}$

( $Wt_{qi}$ : 측정항목 중요도 )

가중평균 오류율 $E = \frac{\sum_{i=1}^{n} e_{wi}}{\sum_{i=1}^{n} Wt_{qi}}$

〈표 4-16〉은 오류율 측정 항목이 추가된 체크리스트의 사례이다.

〈표 4-16〉 오류율 측정이 추가된 체크리스트 사례

| 측정 대상 | 측정 대상 중요도 | 품질 기준 | 품질 기준 중요도 | 세부 기준 | 측정항목 | 측정 항목 중요도 | 측정내용 | 측정 기준 | 측정 방법 | 현 수준 | 측정결과 | | 오류율 | 가중 오류율 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | 총 건수 | 오류 건수 | | |
| △△사 적지 소개동 영상 | 2.4 | 기능성 | 0.4 | 정확성 | 부가요소 정확성(자막 정확성) | 2.5 | **자막은 맞춤법 표기에 따라 작성 되었는가?**<br>5 : 100% 맞춤법 표기 준수<br>4 : 98% 이상 맞춤법 표기 준수<br>3 : 95% 이상 맞춤법 표기 준수<br>2 : 90% 이상 맞춤법 표기 준수<br>1 : 90% 미만 맞춤법 표기 준수 | 100% (콘텐츠 작성표준지침) | Sampling | 4 | 100 | 2 | 2% | 5% |
| | | | | | 부가요소 정확성(사운드 정확성) | 1.6 | **나레이션 시나리오와 사운드 내용은 일치하는가?**<br>5 : 100% 일치<br>4 : 98% 이상 일치<br>3 : 95% 이상 일치<br>2 : 90% 이상 일치<br>1 : 90% 미만 일치 | 100% (콘텐츠 작성표준지침) | Sampling 누락, 잘못읽음, 발음부정확 건수 | 3 | 100 | 4 | 4% | 6.4% |
| | | | | 적절성 | 운용적절성(비디오압축코덱) | 2.5 | **비디오 압축 코덱은 표준을 준수하고 있는가?**<br>5 : 모두 준수<br>4 : 98% 이상 준수<br>3 : 95% 이상 준수<br>2 : 90% 이상 준수<br>1 : 90% 미만 준수 | 100% (콘텐츠 작성표준지침) | 전수 검사 | 5 | 100 | 0 | 0% | 0% |
| | | | | 기능순응성 | 규격화 | 1.9 | **기능성 관련 항목에 대한 표준 지침이 있는가?**<br>5 : 모든 항목에 대해 표준이 문서화되어 있음<br>4 : 일부 항목에 대해 표준이 문서화되어 있음<br>3 : 모든 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>2 : 일부 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>1 : 표준이 없음 | 기능성 관련 항목의 표준이 모두 문서로 명시되어 있어야 함 | 기능성 관련 항목의 표준이 문서로 명시되어 있는지 확인 | 4 | - | - | - | - |
| | | 효율성 | 0.6 | 시간효율성 | 응답속도 | 2.8 | **선택한 동영상이 기준 시간 내에 로딩되는가?**<br>5 : 모두 만족함<br>4 : 90% 이상 만족함<br>3 : 80% 이상 만족함<br>2 : 70% 이상 만족함<br>1 : 70% 미만 만족함 | 〈 1초 | 전수검사 | 3 | 100 | 15 | 15% | 42% |
| | | | | 효율순응성 | 규격화 | 2.5 | **효율성 관련 항목에 대한 표준 지침이 있는가?**<br>5 : 모든 항목에 대해 표준이 문서화되어 있음<br>4 : 일부 항목에 대해 표준이 문서화되어 있음<br>3 : 모든 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>2 : 일부 항목에 대해 표준이 있으나 문서화되어 있지 않음<br>1 : 표준이 없음 | 효율성 관련 항목의 표준이 모두 문서로 명시되어 있어야 함 | 효율성 관련 항목의 표준이 문서로 명시되어 있는지 확인 | 4 | - | - | - | - |

〈표 4-16〉 사례로부터 가중평균 오류율은 다음과 같이 계산된다.

가중평균오류율(E) = (5 + 6.4 + 0 + 42) / (2.5 + 1.6 + 2.5 + 2.8) = 5.7%

# 제4절 오류 원인 분석

## 1. 오류 데이터의 발생 요인

비정형 데이터의 오류는 복합적인 원인으로 발생하는 경우가 많기 때문에 오류 유형을 정확히 판정하는 것은 쉽지 않다. 그러나 메타데이터 부분에 대한 오류는 정형 텍스트 데이터에 대한 데이터 오류 원인 분석 내용과 동일하게 적용할 수 있으며, 비정형 콘텐츠 자체의 경우는 오류 유형에 대한 분석을 거듭하면서 오류 유형을 추가하고 다듬어 특정 조직이나 기업·기관에 적합한 분류 기준을 만드는 것이 바람직하다. 일반적으로 비정형 데이터에 대한 오류가 발생하는 원인에는 다음과 같은 유형이 있다.

### 1.1 작성 표준의 부재

비정형 콘텐츠를 작성하면서 이와 관련한 작성 표준을 먼저 만들어 적용하지 않으면 제작 시 사람마다 또는 콘텐츠마다 제각각의 결과물이 생산될 수 있기 때문에 품질을 유지하기 어려워진다. 메타데이터 항목 구성 및 데이터 작성에 대한 표준을 비롯하여 콘텐츠 작성에 사용되는 코덱이나 화면비율, 자막의 위치와 기본 크기, 기준 환경의 구성 등 많은 요소에 대해 명확한 표준이 준비된 상태에서 구성원 모두가 표준에 수록된 내용을 숙지하고 콘텐츠 작성이 이루어져야 한다.
표준의 부재는 균일하지 않은 결과물의 생산, 잘못된 데이터의 양산, 비규격 데이터의 혼재 등 많은 2차적 오류 상황을

야기 시키는 원인이 되기 때문에 비정형 데이터의 품질문제에 있어서 작성 표준은 매우 중요한 요소가 된다.

### 1.2 부적절한 작성 도구 및 설정 상태

정형 텍스트 데이터의 경우와 달리 비정형 콘텐츠는 대부분 특정 저작 도구를 사용하여 콘텐츠를 제작한다. 이 때문에 어떤 도구를 사용했느냐에 따라 결과물의 품질이 달라질 수도 있고, 작성자가 작성 도구를 얼마나 능숙하게 사용할 수 있느냐에 따라서도 품질이 달라질 수 있으며, 때로는 작성 도구의 기본 설정이 어떠했는지 또는 작성자들 간에 작성 도구의 기본 설정이 일치했는지, 어떠한 값을 사용했는지 등에 따라서도 결과물의 품질이 달라질 수 있다. 이러한 점을 고려하여 저작 도구를 이용한 콘텐츠 제작 시 작성자들이 저작 도구에 익숙해야 함은 물론이고, 도구의 기본 설정 및 사용 값에 대해서도 표준을 정의해야만 기본적인 품질 수준이 확보될 수 있다.

### 1.3 작성 목적에 대한 이해 부족

기본적으로 데이터의 품질을 결정하는 요소 중에서 데이터의 활용 목적은 무시할 수 없는 부분이다. 데이터의 완전성이나 정확성 등 흔히 사용하는 품질기준이 데이터의 활용목적에 따라서 동일한 데이터 상태에 대해 다른 결론을 도출할 수 있기 때문이다. 예를 들어 전화번호는 텔레마케터에게는 매우 중요한 데이터여서 높은 정확도를 요구하지만 급여담당자에게는 그리 중요한 데이터가 아닌 것과 같다. 이와 유사하게 비정형 콘텐츠도 작성 목적이 무엇인지에 따라 그에 맞는

수준의 내용과 품질을 유지하고 있어야 한다. 예를 들어 지역의 사적지를 소개하는 동영상인 경우, 사적지를 소개하면서 대상 사적지의 모습이 너무 작게 잡혀 있거나 또는 영상이 너무 작거나 흐릿하다면 효과가 떨어질 수 있고, 구성 내용이 불충분하거나 실행시간이 너무 짧은 경우에도 목적을 달성하기에 부족할 수 있다. 이와 같이 콘텐츠 작성 목적을 충분히 고려하여 이에 맞는 내용과 품질 수준을 설계하고 이에 대한 작성 표준을 만들어 적용해야만 원하는 목적을 달성할 수 있는 결과물과 품질 수준이 확보될 수 있다. 물론 내용에 대한 판단에 있어서는 평가자의 주관에 다소 영향을 받을 수도 있겠으나 최초 설계 시 이러한 목적이 충분히 고려된 설계가 필요함을 강조하는 것이다.

## 1.4 메타데이터 입력 오류

데이터베이스 내에 텍스트 데이터로 생성된 비정형 콘텐츠의 메타데이터는 정형 텍스트 데이터와 동일한 유형의 오류들이 발생할 수 있다. 즉, 잘못된 값이 입력되거나, 무의미한 값으로 채워지거나, 필요한 데이터가 채워지지 않거나, 또는 동영상과 같은 해당 콘텐츠가 변경되었는데도 관련 메타데이터는 갱신되지 않아 일치하지 않는 정보를 저장하고 있게 되는 경우 등 일반적인 텍스트 데이터에 대해서 발견되는 많은 오류 유형들이 나타날 수 있다. 이에 대해 텍스트 데이터에 대한 오류 방지를 위한 활동들이 그대로 적용될 수 있으며, 특히 메타데이터 구성 항목 중 최초생성일과 같이 변경이 발생하지 않는 항목들과 제목, 내용소개 등과 같이 콘텐츠 변경 시 연동하여 갱신이 필요한 항목들을 구분하여 갱신이 필요한 항목들은 주기적으로 점검을 하는 등의 조치가 필요하다.

## 1.5 메타데이터 불일치

메타데이터는 특정 콘텐츠에 대한 설명을 갖고 있기 때문에 메타데이터를 구성하는 관련 데이터들 간의 일관성이나 정확성, 최신성 등도 중요하지만, 이에 못지않게 해당 콘텐츠와의 일치 여부도 매우 중요하다. 임의의 메타데이터가 해당하는 콘텐츠와 정확하게 연결되는지, 메타데이터의 내용이 해당 콘텐츠에 대한 것과 일치하는지 등 데이터 자체에 대해서만이 아니라 해당 콘텐츠와의 연계성 관점에서도 데이터의 품질이 고려되어야 한다.

## 1.6 품질관리활동 미비

비정형 콘텐츠의 제작 단계에서부터 작성 표준을 기준으로 표준에 대한 부합여부를 관리하고, 주기적인 검사 및 품질보증을 위한 관리 프로세스 확립과 준수 등 품질관리 활동이 필요하며, 이는 제작이 완료되어 운영 중에 있는 콘텐츠에 대해서도 동일하게 적용될 수 있다. 운영 과정에서도 콘텐츠의 추가 제작 및 변경이 일어나고 있고, 생성 후 시간이 많이 경과하여 더 이상 사용되지 않는 콘텐츠의 처리나 메타데이터의 보완 등 다양한 활동에 의해 비정형 데이터에 대한 품질이 영향을 받을 수 있다. 이 때문에 운영 과정에서도 품질관리활동이 지속적으로 필요하게 되며, 이에 대한 품질관리활동이 운영 단계에서도 비정형 콘텐츠에 대해 지속적·체계적으로 이루어지고 있는지에 대한 여부는 매우 중요한 사안이라 할 수 있다. 더불어서 비정형 콘텐츠에 대한 변경관리를 위한 관리 프로세스 등 품질관리 프로세스에 대한 체계화 및 지속적인 수행도 간과해선 안 될 부분이다.

## 2. 오류 원인 분석 방법

정형 텍스트 데이터에 대해 업무규칙을 도출하여 데이터 품질을 측정한 이후 주요 오류 발생 컬럼에 대한 원인분석을 실시하는 것과 마찬가지로, 비정형 콘텐츠에 대해서도 체크리스트를 통해 품질을 측정한 이후 주요 오류 내역을 파악하고 이에 대한 원인분석을 실시한다.

오류 원인 분석 시 메타데이터에 대한 오류인 경우는 정형 텍스트 데이터에 대해서와 같이 오류 발생 컬럼에 대해 오류 발생 유형 · 오류발생 시점 · 오류발생 주기 · 오류가 발생된 데이터의 생성 및 변경 시점을 우선 파악해야 하고, 파악된 오류의 발생 유형 · 오류 발생 시점 · 데이터 생성 시점을 토대로 일시적 혹은 지속적으로 발생되는 데이터인지를 파악하여 오류데이터 발생 추이를 분석한다. 또한 데이터의 오류발생이 데이터 입력 통제나 기타 다른 방법에 의해 억제가 가능한지를 판정한다.

비정형 콘텐츠 자체에 대한 오류인 경우는 오류가 발생한 콘텐츠와 오류가 발생된 측정 항목 및 측정 내용을 토대로 오류 발생한 대상 콘텐츠와 오류발생 유형 · 오류발생 시점 · 오류발생 주기 등을 파악하고, 이를 토대로 지속적으로 발생하는 오류인지, 오류 발생 억제를 위한 방법은 무엇이 있는지 등을 판단한다.

이와 같이 판단된 내용을 근거로 오류가 발생한 대상 콘텐츠의 변경 · 재작성 · 메타데이터 보완 · 표준 보완 등 품질 개선 방안이 도출되어야 한다.

| 집 필 진 | |
|---|---|
| 이 창 한 | 한국데이터베이스진흥원 실장 |
| 김 선 영 | 한국데이터베이스진흥원 팀장 |
| 신 성 수 | 한국데이터베이스진흥원 수석 |
| 서 직 수 | 한국데이터베이스진흥원 수석 |
| 임 성 준 | 한국데이터베이스진흥원 선임연구원 |
| 이 동 훈 | 한국데이터베이스진흥원 연구원 |
| 박 상 용 | 엔코아컨설팅 상무 |
| 명 재 호 | 엔코아컨설팅 상무 |
| 제갈세용 | 엔코아컨설팅 책임컨설턴트 |

데이터 품질관리 시리즈 4

# 데이터 품질진단 절차 및 기법(Ver 1.0)

Data Quality Assessment Procedure Manual

1판 1쇄 인쇄 | 2009년 10월 28일
1판 1쇄 발행 | 2009년 10월 28일

발행인 | 한응수
펴낸곳 | 한국데이터베이스진흥원
주 소 | 서울시 중구 다동 10 한국관광공사 9층
전 화 | 02-3708-5300
팩 스 | 02-318-5040

인 쇄 | (주)드림이노플래너스 02-2276-0811

ISBN 978-89-88474-09-9 93560