공공데이터 표준화를 통한 데이터 품질향상에 관한 연구

김정대[1], 김철롱[2], 임준섭[3], 고광만[4]
[1,2,3,4]상지대학교 컴퓨터공학과
2023015101@sj.sangji.ac.kr, 2023015001@sj.sangji.ac.kr, 2023015002@sj.sangji.ac.kr
kkman@sangji.ac.kr

A study on Data Quality Improvement through the Application of Public Data Standards

Joung-Dae Kim[1], Chul-Rong Kim[2], Joon-Seob Im[3], Kwang-Man Ko[4]
[1,2,3,4]Dept. of Computer Engineering, Sang-Ji University

요 약

공공데이터 품질은 국민의 이용권 보장과 민간에서 활용을 통한 삶의 질을 향상하는데 있어서 중요하다. 현 공공데이터가 운영되는 시스템에서 데이터의 품질 저하가 심각하다. 대부분의 공공데이터는 품질평가 및 개선에 대해서만 논의되고 있다. 구체적인 데이터 표준 적용을 위한 실무 적용 방안에 한계를 갖는다. 본 논문에서는 데이터베이스 설계 시 데이터 표준화 적용 방안과 사례를 제시하였다. 본 연구를 통해 구조적 품질수준이 확보된 데이터베이스 설계 시 데이터 표준화를 수행하고 실무에 기여할 수 있다.

1. 서론

인공지능, 빅데이터, IoT 등 지능 정보기술의 발전으로 데이터의 거래, 유통, 융합 등의 수요가 증가하면서 공공데이터 개방에 대한 민간의 수요가 커지고 있다. 또한 정부는 2013년 "공공데이터의 제공 및 이용 활성화에 관한 법률"을 제정하고, 범정부 공공데이터 포털(data.go.kr)을 통해 양질의 공공데이터를 지속적으로 개방 확대하고 있다. 공공데이터를 이용하는 국민들은 데이터의 양적 확대뿐만 아니라 질적인 개선을 요구하고 있고, 다양한 데이터의 융·복합을 통한 활용 가치를 높일 수 있어, 공공데이터 표준화의 중요성이 더욱 부각되고 있다. 더욱이 기관 업무는 이기종 시스템들이 존재하며, 이는 곧 표준화 되지 않은 데이터로 정확한 정보의 전달 및 공유가 불가능하게 된다. 근본적으로 잘못된 데이터 구조적 설계는 낮은 품질의 데이터를 생산해 낼 수 밖에 없다. 이에 따라 데이터 구조의 품질수준을 높이기 위한 연구가 필요하다.

2. 관련 연구

2-1. 데이터 품질 개요

데이터 품질(Data Quality)은 데이터에 대한 기관과 사용자의 만족도를 충족시키는 것이라고 정의할 수 있다. 데이터 품질은 완전성, 유일성, 유효성, 일관성, 정확성 등 5개 일반적으로 데이터 품질기준을 활용하여 정의한다[1].

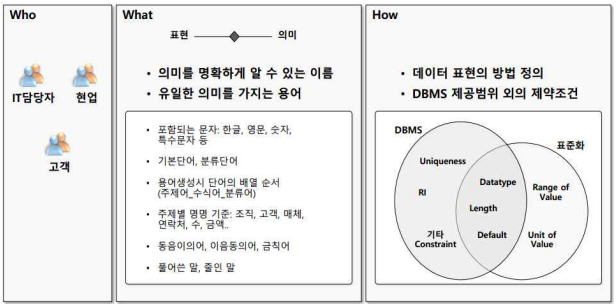
<표 1> 데이터 품질 기준 정의

품질기준	정의
완전성 (Completeness)	필수항목에 누락이 없어야 한다.
유일성 (Uniqueness)	데이터 항목은 유일해야하며 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 반영되어야 한다.

2-2. 데이터 표준화

데이터 표준화의 정의는 정보시스템을 사용하는 모든 사용자(Who)가 동일한 데이터를 같은 의미(What)로 해석하고, 동일한 방법(How)으로 접근하고 사용할 수 있는 원칙과 기준을 합의하고 관리하

는 활동이다.



(그림 1) 데이터 표준화

이는 데이터의 일관성 확보를 통해 데이터의 품질 수준을 높이는 방법으로 데이터의 품질을 높이는 가장 현실적인 방안이다. 데이터 표준화 대상으로는 단어, 용어, 도메인, 코드로 정의한다. 데이터 품질관리는 기관에서 혹은 조직 내외부에서 사용하는 정보시스템 및 데이터베이스의 데이터를 지속적으로 관리하고 개선하는 활동을 의미한다.

3. 데이터 표준화 구현 방안

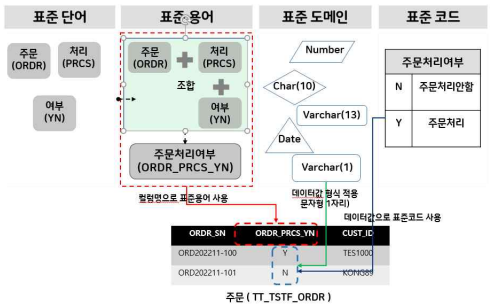
3-1. 데이터 표준화 대상 및 구성

데이터 품질관리 수준을 높이기 위해서 구체적인 적용 방안으로는 기존 단위 시스템에서 사용하는 데이터를 활용하여 표준 데이터를 구축한 후 시스템 고도화 및 신규 구축 시스템을 대상으로 데이터 표준화를 적용한다. 본 논문에서는 데이터 표준화의 필요성과 데이터 표준화 절차 및 지침을 제시한다.

<표 2> 데이터 표준화 대상

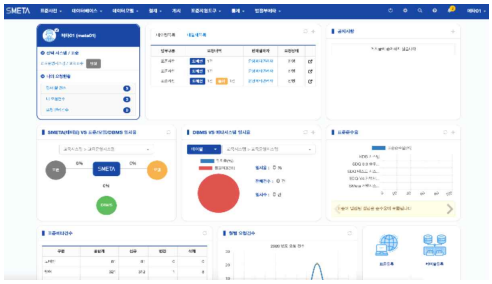
표준화 대상	설명
표준 단어	일정한 뜻을 가지는 말의 최소 단위
표준 용어	업무 및 일상에서 사용되는 표준단어의 조합
표준 도메인	표준용어의 데이터 입력 형식 정의
표준 코드	데이터의 값을 정형화하고, 공통으로 사용되는 기호체계

생성된 표준 용어, 표준 코드, 표준 도메인은 데이터베이스 설계 시 테이블, 컬럼, 데이터 값, 코멘트 등으로 사용된다.



(그림 2) 데이터 표준 사용

3-2. 데이터 표준화를 위한 관리도구

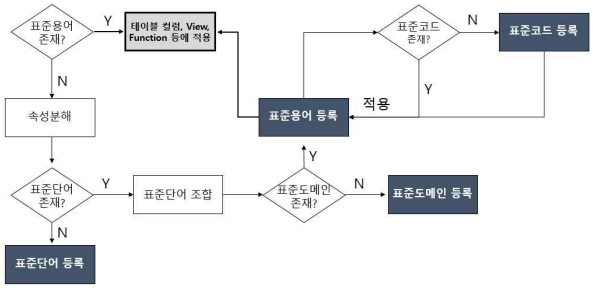


(그림 3) 표준메타관리 도구

표준 메타관리 솔루션을 활용하여, 표준 데이터(단어, 용어, 도메인, 코드 등)를 관리한다. 표준 데이터의 신규 등록 및 변경 시 관리 프로세스를 통해 진행된다.

3-3. 데이터 표준화를 위한 관리 프로세스

데이터 표준화 관리 프로세스는 논리모델, 물리모델 설계 시 적용되며, 또한 관리자・담당자 등의 역할과 요청・검토・기각・승인・확인 등의 절차가 필요하다.



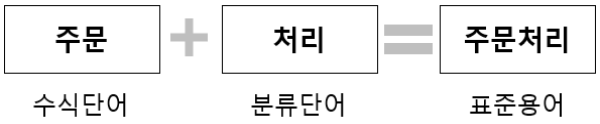
(그림 4) 데이터 표준 관리 프로세스

3-4. 데이터 표준화를 위한 설계 지침

(1) 표준단어 설계

- 수식단어, 분류단어로 구분
- 수식어는 분류단어를 제외한 나머지 단어로 표준

용어 구성 시 분류단어 앞에 위치.
- 분류어는 표준 용어 구성 시 마지막에 위치하며, 표준 용어가 가질 수 있는 데이터의 형식과 범위를 한정.



(그림 5) 표준단어 분류 및 표준용어 구성

- 표준단어는 명사형으로 정의한다.

(2) 표준 도메인 설계
- 표준 용어가 가질 수 있는 데이터 타입, 길이를 제한하는 값으로 설정.
- 표준 도메인은 도메인명, 도메인영문명, 도메인유형, 길이, 타입 등으로 구성.

(3) 표준용어 설계
- 표준단어와 표준 도메인은 최종적으로 표준 용어를 구성.
- 표준용어는 두 개 이상의 표준단어로 구성하며, 마지막 표준단어는 분류단어로 구성.(그림 5 참조)
- 단어와 단어 사이에는 '_'(언더스코어)로 연결하여 구성.
- 표준용어는 하나의 도메인과 결합
- 표준용어는 용어명, 용어영문명, 도메인명, 도메인그룹, 설명 등으로 구성.

(4) 표준코드 설계
- 표준코드는 도메인그룹명, 도메인명, 코드, 코드명 등으로 구성한다.

4. 데이터표준화 구현 사례 및 효과검증

4-1. 데이터 표준화 구현 사례

데이터 표준화를 구현한 사례를 중심으로 데이터표준화가 수행된 현황을 제시한다.

<표 3> 데이터 표준 사전 구축 현황

구 분	구 축 건 수
표준 단어	3,302건
표준 도메인	105건
표준 용어	7,106건
표준 코드	1,885건

(그림6)과 같이 표준단어는, 단어명, 영문명, 영문약어명, 설명 등으로 정의된다.

(그림 6) 표준단어 등록 화면

(그림7)과 같이 표준용어는, 용어명, 용어영문명, 타입, 길이 및 도메인이 결합되어 정의된다.

(그림 7) 표준용어 등록 화면

4-2. 데이터 표준화 효과성 검증

(1) 정량적 효과

<표 4> 데이터 표준적용률

시스템	전체 컬럼수	표준적용 컬럼수	표준 적용률(%)
A시스템	51	1	1.96
B시스템	256	4	1.5
C시스템	**27**	**27**	**100**

(표4)는 A, B시스템은 데이터베이스 표준화가 적용되지 않은 시스템의 표준 적용률을 측정한 결과이며, C시스템은 본 논문의 표준화 구현방안이 적용된 결과이다. 표준화를 적용한 시스템이 그렇지 않은 시스템보다 데이터의 구조적 일관성과 품질특성이 우수함을 알 수 있다. 데이터 표준 적용률(%)은 아래 공식을 활용하여 산정하였다[2].

$$\frac{\text{진단·평가대상 } DB\text{에 데이터표준이 적용된 컬럼 수}}{\text{진단·평가대상 } DB\text{전체 컬럼 수}} * 100$$

(2) 정성적 효과

데이터베이스 설계 수행 시 본 논문의 구현방안을 적용하여 정량적 효과인 구조적인 품질향상 외에 다음과 같은 정성적 효과를 얻게 되었다.

- 표준화된 테이블 설계를 위한 데이터 표준화 가이드, 관리 규칙을 제공하여 설계 생산성 향상
- DB구조 변경에 대한 영향도 분석용이 및 운영편의성 향상됨.
- 표준메타관리 도구를 통한 데이터베이스 설계와 관련한 담당자들 간의 의사소통 편의성 제공

5. 결론

공공데이터 표준화를 통해 공공데이터에 대한 구조적 품질향상을 확보할 수 있으며, 이를 통해 국민 및 기업에 신뢰성 있는 표준화된 데이터를 제공할 수 있다. 본 논문에서는 데이터 표준화에 대한 개념과 데이터 표준 구현 방안에 대해서 제시하였다. 그 결과 데이터 표준화를 수행하지 않은 경우와 수행한 경우를 비교했을 때 일관성 및 신뢰성 측면에 있어서 데이터 표준화를 수행한 경우 구조적 품질수준이 보장됨을 확인 할 수 있다. 향후에는 XML, HTML 등의 반정형 데이터, 텍스트, 이미지, 동영상 등의 비정형 데이터에 대한 표준화 적용 방안에 대한 연구가 필요하다.

감사의 글: 이 성과는 과학기술정보통신부의 재원으로 정보통신산업진흥원의 지원을 받아 수행된 연구임(디지털트윈 융합 의료혁신 선도 사업).

참고문헌

[1] 한국데이터베이스진흥원 “데이터 품질진단 절차 및 기법(Ver1.0)”, 2009.11

[2] NIA 한국지능정보사회진흥원 “2023년 공공데이터 품질관리 수준진단・평가 매뉴얼”, 2023.4