데이터표준화 사례를 통한 데이터 품질 향상에 대한 연구

김진섭°
고려대학교 컴퓨터정보통신대학원
jinseob.kim@samsung.com

A Research of Enhancing Data Quality Through the Implementation of Data Standardization

Kim Jinseob°
Graduate School of Computer and Information Technology, Korea University

요 약

데이터품질이 기업의 경쟁력에 영향을 주는 핵심요소임에도 불구하고, 현 정보시스템 현실에서는 데이터품질 저하라는 심각한 상황을 맞고 있다. 데이터품질을 개선시키기 위한 여러 가지 방안들이 논의되고 있지만, 대부분 현상 데이터에 대한 품질 평가 및 개선에 한정되거나, 개선방안의 구체성이 부족하여 실무적 적용에 한계를 갖는다. 본 연구에서는 데이터표준화 개념을 데이터베이스 설계와 병행하여 수행할 수 있도록 구체적인 구현방안과 사례를 제시하였다. 데이터표준화는 각 단위시스템의 데이터에 대한 명칭 및 도메인에 대한 표준원칙을 수립하여 표준데이터를 구축한 후 전체 시스템에 적용하는 방법이다. 본 연구의 구현방안은 표준데이터 구축이 선행되지 않은 경우에도 데이터의 구조적 품질수준이 보장된 데이터베이스 설계를 수행하고자 하는 실무에 기여할 수 있다.

1. 서 론

오늘날 기업의 정보화 및 고도화 요구가 지속적으로 증가되고, 정보제공기반의 다각화가 심화됨에 따라 데이터의 깊이와 양은 급속하게 증가되고 있다. 또한, 정보의 가치에 대한 인식 변화에 따라 이용자들은 점차 정보서비스의 품질, 가치 등에 민감해지고 있다.[1]

이에 따라 점차 증대되는 대용량 데이터의 품질은 점차 기업 경쟁력 확보를 위한 이슈로 등장하고 있다. 그러함에도 불구하고 대부분의 기업은 Poor Data Quality로 인해 많은 어려움을 갖는다. 데이터품질이 기업의 경쟁력에 영향을 주는 핵심요소임에도 불구하고, 현 정보시스템 현실에서는 데이터품질 저하라는 심각한 상황을 맞고 있다. 데이터의 품질저하로 인한 비용이 전체 제조업의 내부 실패 비용(손실 및 재작업 비용)의 40~60%를 차지하고 있다는 연구결과까지 보고되고 있다.[2]

이제 데이터의 품질은 기업의 전 측면에 걸쳐 중요한 영향요인으로써 작용하고 있으며, 정보화 시대의 기업에 있어 중요한 경쟁력의 척도로써 작용하고 있다. 이에 따라 데이터품질관리에 대한 연구들이 활발하게 이루어지고 있으며, 실제 실무에서도 관련 Tool들이 일부 활용되고 있는 추세이다.

이처럼 데이터의 품질의 중요성을 고려할 때, 데이터 구조 설계가 가지는 책임을 간과하지 않아야 한다. 근본적으로 잘못된 데이터 구조설계에 의해 만들어진 데이터는 비록 그 데이터 값 자체는 실물과 일치되어 입력된다 하더라도 품질이 좋은 데이터라고 할 수 없다. 데이터 구조의 근본적인 문제점을 가진 데이터들은 지속적으로 품질이 낮은 데이터를 생산해 낼 수 밖에 없다는 한계성을 갖게 되기 때문이다.[3]

이에 따라 데이터의 구조적 품질수준을 높이기 위한 연구가 필요하다. 본 연구에서는 데이터표준화 개념을 적용하여 데이터의 구조적 품질수준을 높일 수 있는 방안을 제시하였다. 제시된 방안은 데이터베이스 설계와 병행하여 표준데이터를 구축하고 이를 통해 자연스럽게 데이터표준화 작업이 수행될 수 있도록 함으로써, 데이터베이스 품질 뿐만 아니라 설계의 생산성을 향상시킬 수 있는 실무적인 적용 기반을 마련하였다.

본 논문의 구성은 다음과 같다. 2장에서 데이터품질 에 대한 개념과 관련 연구에 대해 기술한다. 3장에서 데이터베이스 설계와 병행하여 데이터표준화 작업을 수행하기 위한 방안을 제시한다. 4장에서는 제시된 방안을 활용한 실제 구축사례를 보이고, 구축결과를 분석하여 제시된 방안의 효율성에 대해 입증한다. 5장의 결론으로 본 논문을 마무리한다.

2. 관련 연구

2.1 데이터품질의 개념과 특성

Larry P. English는 데이터품질(Data Quality)의 정의를 기업과 고객의 목표를 달성하기 위해 데이터에 대한 이해관계자의 기대를 충족시키는 것이라고 정의하였다. 또한, 데이터품질 유지를 위한 원칙으로 과학적 기법을 통해 고객에 집중하여 데이터에 대한 개선활동을 수행하도록 제시하고 있다.[2]

데이터 품질 특성(Data Quality Characteristics)은 소프트웨어 품질 특성[4]과는 달리 표준이 명확히 정립되어 있지 않고, 각기 필요성에 따라 조금씩 연구가 진행되어 왔다. 그 중 대표적인 것으로 Wang의 연구[5]를 들 수 있는데, 데이터품질은 4가지 차원, 즉 정확성(accuracy), 적시성(timeless), 완전성(completeness), 일관성(Consistency)으로 구분된다.

각 데이터 품질 특성에 대한 측정은 소프트웨어 품질 측정 표준인 ISO/IEC 9126을 기반으로 데이터 품질을 측정하기 위한 방안에 대한 연구가 진행되고 있다. 최병주[6]는 오류데이터를 분류하고, 그것으로부터 데이터 품질 특성을 측정하기 위한 메트릭을 제시하였다.

2.2 데이터표준화

데이터표준화[7]는 단위 시스템별로 산재되어 있는 데이터에 대하여 명칭 및 도메인에 대한 표준 원칙을 수립하여 표준데이터를 구축한 후 전체적인 시스템에 적용하는 방법으로, 데이터

의 품질 특성을 높일 수 있는 현실적인 방안이다. 표준데이터란 정보시스템에서 사용하는 용어, 도메인, 코드 및 기타 데이터 관련요소에 대해 공통된 형식과 내용으로 정의하여 사용하는 표준 관련 데이터를 의미한다.

미 국방부에서는 통합되고 효율적인 방식으로 미션을 수행하기 위한 DoD 8320.1-M-1 'Data Element Standardization Procedures' [8]를 발표하였다. DoD 8320.1에서는 표준데이터를 개발, 승인, 구현, 유지보수 하는 절차 및 데이터 관리정책을 제시하였다.

배창호[9]는 국방데이터베이스 통합을 위한 데이터표준화 방안을 제안하였다. 이 연구에서는 데이터표준화의 필요성과 데이터표준화 절차 및 지침이 제시되어 있다.

박주석[3]은 데이터의 구조적 품질관리 성숙도 모델을 제시하였다. 이 모델은 Level 1에서 Level 4 까지의 4단계로 구성되어 있으며, 상위 단계로 올라갈수록 데이터의 구조적 품질관리 수준이 성숙된다고 정의하였다. 이 연구에서는 표준데이터를 先 구축한 후 신규 시스템 개발 시 참조하도록 한 국방표준 데이터 관리시스템 구축사례가 제시되어 있다.

2.3 데이터품질관리

데이터품질관리[7]란 기관이나 조직 내외부의 정보시스템 및 DB 사용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 및 개선활동을 의미한다. 데이터품질관리에 대한 초기 연구는 품질 측정에 대한 현상분석 중심이었으나, 점차적으로 품질개선을 위한 모델 중심으로 바뀌고 있다.

MIT의 TDQM(Total Data Quality Management) 프로그램[5]은 데이터품질을 본질적(Intrinsic) 품질, 연관적(Contextual) 품질, 표현적(Representational) 품질 등의 카테고리로 분류하고, 각 카테고리 별로 데이터품질 문제가 발생하는 패턴 및 개선방안을 연구하고 있다.

Larry P. English의 TIQM(Total Information Quality Management) 모델[2]은 6 단계의 프로세스로 구성되어 있으며, 각 단계의 프로세스들은 평가, 유지보수, 데이터 이행 통제, 유지보수를 위한 프로세스 개선 및 조직을 데이터 품질의 문화로 전환하는 프로세스로 구성된다.

국내 데이터베이스진흥센터에서는 데이터 품질관리지침[7]을 통해 품질개선을 위한 프레임워크를 제시하였다. 데이터품질관리 프레임워크는 Enterprise Architecture의 개념을 도입한 것으로 표준데이터를 정보시스템의 데이터 품질 확보를 위한 필수 요소로 정의하였다.

3. 데이터표준화 구현방안

3.1 데이터표준화에 대한 연구 방향 및 범위

데이터품질에 대한 2장에서의 선행 연구들을 종합해 보면, 데이터품질을 향상시키기 위한 여러 가지 개선방안들이 제시되고 있음을 알 수 있다. 최근의 연구들은 평가나 관리프로세스 측면에 중점을 두는데, 이는 제품의 품질을 향상시키기 위해서는 프로세스 품질 향상이 선행되어야 한다는 소프트웨어프로세스 평가모델인 CMMI(Capability Maturity Model Integration) [10]의 사상과도 부합된다.

그러나, 기존의 데이터 품질개선에 대한 연구는 현상적인 데이터의 품질 개선에 한정되거나, 개선방안의 구체성이 부족한 측면이 많아 실무적 적용에 한계를 가지고 있다. 이에 따라 실제 실무에 적용하여 데이터품질관리 수준을 높이기 위해서는 보다 구체적인 방안이 제시될 필요가 있다. 또한 기존 연구의 데이터표준화는 AS-IS 데이터를 이용하여 전사적인 Top-Down 방식으로 표준데이터를 구축한 후 각 단위시스템에 적용하는 방식이어서 데이터베이스 통합 또는 차세대 프로젝트 등에 적용하기에는 유리하나, AS-IS 데이터가 없거나 프로젝트 초반부터 표준데이터 구축을 위한 자원을 확보하기가 어려운 보통의 실무에 적용하기에는 많은 어려움을 갖게 된다.

이에 따라 본 연구는 기존 연구의 데이터표준화 개념을 적용하되 절차를 개선하여 표준데이터 구축을 데이터베이스 설계와 병행하여 수행할 수 있는 구체적인 구현방안과 사례를 제시하였고, 데이터베이스 품질 뿐만 아니라 설계의 생산성을 향상시킬 수 있는 실무적인 적용 기반을 마련하였다.

본 연구에서의 <표 1>과 같이 표준데이터를 기본으로 논리속성과 물리칼럼까지 데이터표준화 범위에 포함시켰다.

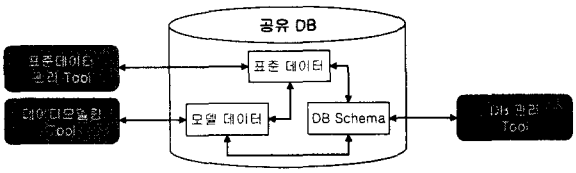
<표 1> 데이터표준화 구성요소

분류	구분	설명
표준 데이터	표준단어	표준용어를 구성하는 의미의 최소단위
	표준도메인	데이터 형식의 일관성을 갖는 그룹
	표준용어	독립적이고 구체적인 의미를 갖는 표준단어들의 조합
	표준코드	데이터 값을 정형화하기 위한 기호
모델 데이터	논리속성	엔티티의 속성(Attribute)을 의미
	물리칼럼	테이블의 칼럼(Column)을 의미

3.2 데이터표준화를 위한 환경 및 관리프로세스

(1) 데이터표준화를 위한 환경

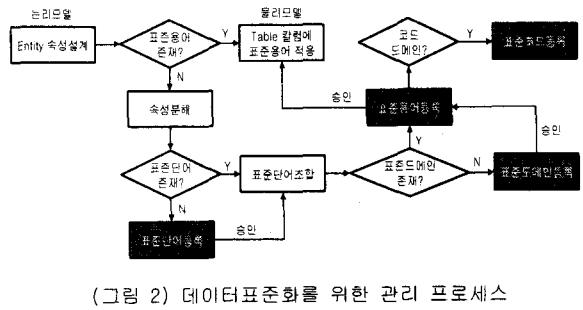
데이터베이스 설계와 병행하여 표준데이터를 구축하기 위해서는 설계정보와 표준데이터정보가 서로 공유될 수 있도록 하여야 한다. (그림 1)에서 제시된 것처럼 각 메타정보에 대한 Repository를 공유할 수 있는 환경이 필요하다.



(그림1) 데이터표준화를 위한 환경

(2) 데이터표준화를 위한 관리 프로세스

데이터베이스 설계가 데이터표준화와 병행하기 위해서는 엄격한 관리 프로세스 유지가 필요하다. 또한 프로세스 유지를 위한 DA(Data Architect)의 역량과 설계자들의 준수노력이 함께 요구된다. (그림2)에서 제시된 것처럼 논리 엔티티의 속성 설계 시작부터 표준데이터가 발생할 수 있도록 한다.



(그림 2) 데이터표준화를 위한 관리 프로세스

3.3 데이터표준화를 위한 설계지침

(1) 표준단어 설계

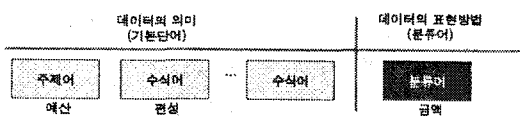
- O 기본단어와 분류단어로 구분한다.
- O 분류단어는 데이터의 표현방법을 나타내므로 불분명한 저장형태는 배제한다. (배제 예: 의견, 전화, 결과 등)
- O 명사형으로 정의하고, 특수기호를 포함하지 않는다.
- O 표준어 및 업무적으로 사용되는 단어를 사용한다.
- O 이음동의어는 대표 단어를 선정하여 사용한다.
- O 하나의 한글표준단어에 하나의 영문약어를 정의한다.
- O 모든 표준단어에 단어가 의미하는 내용을 정의한다.

(2) 표준도메인 설계

- O 도메인명은 데이터의 타입 및 속성을 인식할 수 있는 용어로 선정한다. (예: 번호, 일자, 코드, 수, 명 등)
- O 번호와 코드 도메인은 표준용어별로 각각의 하위 도메인을 생성할 수 있다.
  (예: 주민등록번호, 접수구분코드 등)
- O 금액이나 수 성격의 도메인은 데이터 길이 유형별로 각각의 하위 도메인을 생성할 수 있다.
  (예: 금액10, 금액13, 수3, 수9 등)
- O 하나의 표준도메인에 하나의 데이터 타입 및 길이를 부여한다.

(3) 표준용어 설계

- O 표준용어는 두 개 이상의 표준단어로 구성된다.
- O 표준용어는 (그림 3)처럼 주제어, 수식어, 분류어를 조합하여 만든다. 용어의 맨 우측은 데이터의 표현방법을 나타내는 분류단어(분류어)이어야 한다.
  - 주제어의 예: 예산, 수입, 지출 등
  - 수식어의 예: 보류, 승인, 지급, 신규, 등록 등
  - 분류어의 예: 기간, 액, 수, 명, 값 등
- O 표준용어의 영문명은 표준단어에 적용된 영문약어를 '_' 를 이용하여 조합한다.
- O 하나의 표준용어에 하나의 도메인을 결합한다.
- O 하나의 표준용어는 유일해야 한다.
- O 용어가 의미하는 내용을 기술하여, 동일한 의미를 갖는 용어들은 하나의 표준용어로 정의한다.
  (예: 주민번호, 주민등록번호 -> 주민등록번호)



(그림 3) 표준용어의 구성요소

(4) 표준코드 설계

- O 코드도메인이 적용된 표준용어에 대해서는 표준코드를 정의한다.
- O 정의된 표준코드는 코드 설계서에 기술한다.

(5) 논리 엔티티 속성 및 물리 테이블 칼럼 설계

- O 논리 엔티티 속성은 표준용어를 그대로 적용한다.
- O 물리 테이블 칼럼은 적용된 표준용어의 영문명을 그대로 사용한다.
- O 설계 예: '예산편성금액' 인 속성을 설계
  - 표준단어 : 예산(BUGT), 편성(COMP), 금액(AMT)
  - 표준도메인 : 금액13 -> NUMBER(13)
  - 표준용어 : 예산편성금액 + 금액13
  - 테이블 칼럼 : BUGT_COMP_AMT NUMBER(13)

4. 데이터표준화 구현사례 및 검증

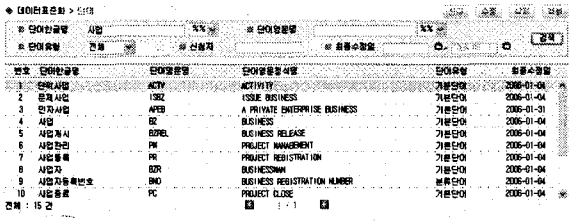
4.1 데이터표준화 구현사례

본 장에서는 데이터표준화를 구현한 사례를 중심으로 기술한다. 데이터표준화가 수행된 프로젝트의 설계단계 종료 시점까지 구축된 표준 및 모델 데이터 현황을 <표 2>에 제시하였다.

<표 2> 표준 및 모델 데이터 구축 현황

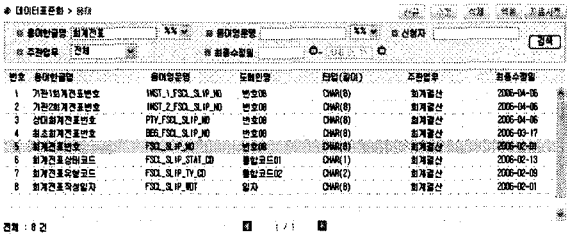
데이터 구분	상세 구분	구축 건수
표준 데이터	표준 단어	787
	표준 도메인	72
	표준 용어	1901
모델 데이터	엔티티 / 테이블	440
	속성 / 칼럼	7442

(그림 4)는 표준단어를 등록하는 화면이다. 각 표준단어별로 영문약어 및 단어유형이 정의되어 있는 것을 알 수 있다.



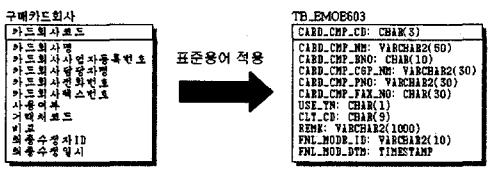
(그림 4) 표준단어 등록 화면

(그림 5)는 표준용어를 등록하는 화면이다. 각 표준용어별로 영문명 및 도메인이 결합되어 정의되어 있는 것을 알 수 있다.



(그림 5) 표준용어 등록 화면

(그림 6)은 표준용어를 적용한 데이터모델 예이다. 좌측은 논리모델이고, 우측은 물리모델이다. 논리 속성에 대응하는 물리 칼럼은 표준용어를 적용하여 별도 작업 없이 자동 반영된다.



(그림 6) 데이터모델

4.2 효율성 검증

(1) 데이터품질 측정 메트릭

품질 측정을 위한 메트릭은 전체 데이터 구조의 일관성 측면에서 오류가 발생한 비율을 이용하여 구하도록 정의한 메트릭[6]을 적용하였으며, 식(1)과 같다.

각 칼럼별 오류 데이터(c_{ij})는 표준데이터와 일관성이 맞지 않는 경우 오류로 판단하였으며, 가중치(w_j)는 각 칼럼의 개수가 전체 칼럼 개수에서 차지하는 비율로 계산하였다.

<정의> 데이터 품질 측정 메트릭 Q

- Q : 데이터 품질 측정값
- N : 총 품질 측정 대상 속성 개수
- T : 칼럼별 가중치를 부여한 총 오류 속성 개수
- m_i : 총 칼럼 수
- c_{ij} : 각 칼럼에 해당하는 오류 속성의 개수
- w_j : 각 칼럼에 해당하는 가중치

$$Q = 1 - \frac{T}{N} \qquad \text{(식 1)}$$

$$T = \sum_{j=1}^{m_i} c_{ij} \times w_j$$

(2) 정량적 효과

본 논문에 제시된 구현방안을 적용하여 데이터표준화가 수행된 설계결과를 평가하기 위해, 규모 및 일정이 유사한 데이터표준화를 수행하지 않은 他 프로젝트와 (식 1)을 적용하여 일관성에 대한 측정결과를 비교하였다. <표 3>에 제시된 결과에 따르면 본 논문의 구현방안을 적용한 프로젝트가 그렇지 않은 경우보다 일관성에 대한 데이터의 구조적 품질특성이 우수함을 보여준다.

<표 3> 프로젝트 간 비교

구분	사례 프로젝트	他 프로젝트
테이블 수	440	367
N	7442	3103
T	0	0.086
Q	0	0.086

<표 4>는 데이터표준화를 데이터베이스 설계와 병행하여 수행한 결과 설계가 진행되어 N값이 증가하더라도 데이터의 구조적 품질특성이 거의 유사한 결과를 보여준다. 이는 표준데이터의 확보가 선행되지 않은 경우에도 표준데이터의 구축과 데이터베이스 설계를 병행하는 경우 데이터품질 수준의 적정성이 확보됨을 나타낸다.

<표 4> 진행단계 별 비교

구분	2월	4월	6월
테이블 수	219	372	440
N	4067	6238	7442
T	0	2	0
Q	0	0	0

(3) 정성적 효과

본 논문의 구현방안을 적용하여 데이터베이스 설계를 수행한 결과 정량적 효과인 구조적 품질 향상 측면 외에 다음과 같은 정성적 효과를 갖게 되었다.

- ○ 논리 데이터모델링과 동시에 표준화된 테이블 설계가 자동적으로 수행되어 설계 생산성이 향상됨
- ○ 표준화된 칼럼을 통해 DB 구조의 변경에 대한 영향도를 쉽게 파악할 수 있게 되어 운영편의성을 제공함
- ○ 구축된 표준용어사전을 프로그램의 용어 표준에 활용하여 개발표준화를 유도하고, 데이터 컨버전의 기초자료로 활용함
- ○ 표준화된 의사소통 수단을 통해 데이터의 연계, 통합 및 EUC(End-User-Computing)를 실현할 때 관리의 편의성을 제공함

5. 결론 및 향후 과제

데이터품질은 기업의 경쟁력 확보를 위한 중요한 영향요인으로 데이터표준화를 통해 데이터의 구조적 품질수준을 보장할 필요가 있다.

본 논문에서는 실무에서 데이터표준화 개념을 데이터베이스 설계와 병행하여 수행할 수 있게 함으로써 데이터의 구조적 품질과 설계생산성을 향상시킬 수 있도록 하는 방안을 제시하였다. 구현방안으로는 데이터표준화를 위한 환경과 관리프로세스 및 설계지침을 기술하였다. 본 논문에서 제시한 구현방안의 효율성을 측정하기 위해 데이터베이스 속성 설계의 일관성에 대한 메트릭을 적용하였으며, 그 결과 데이터표준화를 수행하지 않은 사례에 비해 일관성 측면에서 데이터의 구조적 품질특성이 우수한 결과를 보였다. 본 논문에서 제시된 구현방안은 표준데이터 구축이 선행되지 않은 경우에도 구조적 품질수준이 보장된 데이터베이스 설계를 수행하고자 하는 실무에 기여할 수 있다.

향후에는 논리모델의 Entity 속성을 분해할 때 형태소 단위로 이루어진 한국어의 특성을 고려하여 표준데이터 구축 시 설계속성의 자동분해 기법에 대한 연구가 필요하다.

참고문헌

[1] 한국데이터베이스진흥센터, "2006 데이터베이스백서", 2006.6

[2] Larry P. English, "Improving Data Warehouse and Business Information Quality", Wiley, 1999.2

[3] 김찬수,박주석, "데이터 구조적 분석을 통한 성숙도 모델 개발", 경희대학교, 2004.2

[4] ISO/IEC 9126-1,2,3, JTC 1 SC 7 WG 6(Evaluation & Metrics) Documents, 1996.11

[5] Diane M. Strong, Yang W. Lee, and Richard Y. Wang, "Data Quality in Context", Communications of the ACM, Vol.40 No.5, 1997.5

[6] 양자영,최병주, "데이터품질 측정도구", 한국정보과학회 논문지, 컴퓨팅의 실제 제 9권 제 3호, 2003.6

[7] 한국데이터베이스진흥센터, "데이터 품질관리 지침(Ver. 2.0)", 2005.11

[8] DoD 4000.25-13-M, "DoD Logistics Data Element Standardization and Management Program Procedures", 1996.6

[9] 배창호, "국방데이터베이스 통합을 위한 데이터표준화 방안", 숭실대학교, 2000.12

[10] Margaret K. Kulpa, Kent A. Johnson, "Interpreting the CMMI", AUERBACH, 2003.4