정보 구조 그래프를 이용한 통합 데이터 품질 관리 방안 연구

이춘열*

An Implementation of Total Data Quality Management Using an Information Structure Graph

Choon Yeul Lee

Abstract

This study presents a database quality evaluation framework. As a way to build a framework, this study expands data quality management to include data transformation processes as well as data. Further, an information structure graph is applied to represent data transformations processes.

An information structure graph is absed on a relational database scheme. Thus, data transformation processes may be stored in a relational database. This kind of integration of data transformation metadata with technical metadata eases evaluation of database qualities and their causes.

Keywords: database evaluation, data management, metadata, data evaluation framework

논문접수일 : 2003년 8월 25일 논문게재확정일 : 2003년 10월 21일

* 국민대학교 비즈니스IT전문대학원 부교수, cylee@kookmin.ac.kr

1. 서론

현재 국내 기업들은 정보화를 위하여 많은 자원을 투입하고 있으며, 이에 대한 관심 또한 증대하고 있다. 특히 근자에는 단편적인 정보 시스템의 구현에서 이들 사이의 정보 교환 및 공유가 강조되고 있으며, 단순 데이터 처리보다는 보다 정교한 정보 및 지식의 도출이 주요 관심사로 대두하고 있다.

정보 공유의 확산과 데이터 이용 기법의 고도화는 보다 정교한 데이터 관리를 요구하게 된다. 즉 보다 많은 사용자들이 데이터를 공유하게 됨으로써 데이터의 내용이나 의미, 출처 등에 친근하지 않은 일반 사용자들이 데이터를 사용하게 될 기회가 늘어나게 된다. 이 결과 정보의 오류가 사용자들에게 여과 없이 직접 영향을 미치게 되었다. 부연하면, 정보 이용자들이 스스로 데이터 오류를 식별하기가 쉽지 않다. 기존의 단일 시스템 위주의 데이터관리와 비교하여 데이터의 오류와 같은 품질저하가 더욱 심각한 영향을 끼치게 되었다. 즉, 데이터 오류로 인하여 다양한 형태의 파급 효과들이 발생하고 있으며, 상호 연관된 응용 시스템들은 이러한 데이터 오류의 확산을 가속화 시키게 되었다. 이는 데이터 품질의 문제가 다양한 원천으로부터 데이터가 수집되는 대형 데이터베이스에서 더 흔하게 발생하고 있다는 연구 결과들로부터도 확인할 수 있다.

데이터 품질의 향상을 위하여서는 보다 철저한 데이터의 검증 등 많은 노력이 경주되어 왔다. 그리고 데이터의 불일치나 오류를 방지할 수 있는 데이터베이스 구조가 데이터베이스 설계의 주요 이슈로 취급되어 왔다. 그리고 최근에는 데이터 관리가 데이터의 검증이나 데이터베이스의 설계 등과 같은 일부 단계에서만 이루어지는 것이 아니라 데이터 수명 주기 전체에 걸쳐 지속적으로 이루어지는 통합 데이터 품질 관리 (Total Data Quality Management)의 필요성이 인식되고 있다. 이는 마치 생산 활동에서 제품의 품질관리가 완성된 제품의 품질 측정이나 불량품의 제거로부터 생산 공정을 향상시키며, 고품질의 제품을 설계하는 통합 품질관리 (Total Quality Management)로 확장되었던 것과 비슷한 과정이라고 할 수 있다.

데이터의 품질 또한 초기에는 데이터 자체의 품질에만 관심을 기울였으나, 차츰 데이터의 생성 과정, 데이터베이스 스키마 등에 보다 많은 관심이 기울이고 있다. 그러나 현실적으로 데이터 관리는 정보 시스템 개발 과정에서 데이터베이스 설계의 일부로서 이루어져왔다. 즉, 데이터베이스 스키마를 중심으로 데이터에 대한 명세를 저장·관리하는 수준에서 이루어지고 있다. 따라서 종합적 데이터품질관리가 이루어지기 위하여서는 데이터베이스 스키마를 위주로 한 메타 데이터 수준을 넘어서는 보다 확장된 데이터 관리 모형이 필요하며, 이에 근거한 데이터 관리 환경이 구축되어야 한다.

이를 위하여 본 연구에서는, 데이터 자체의 품질 뿐만이 아니라 데이터 품질에 영향을 미치는 모든 요소들을 포함하는 통합 데이타 품질 모형을 제시하고, 제시된 모형을 기반으로 효과적인 품질 관리를 통하여 높은 수준의 데이터 품질을 유지하기 위한 메타 데이터 관리 방안을 제시한다. 제시된 데이터 품질 관리 모형은 정보구조그래프 (Information Structure Graph)를 이용하여 생성 프로세스에 대한 메타 데이터를 표현하며, 이를 이용하여 품질 데이터를 관리하고 활용하는 방안을 제시한다. 이하 제2장에서는 통합 데이터 품질 평가 모형을 제시하고, 제3장에서는 정보구조그래프를 이용한 메타 데이터 모형을 제시한다. 그리고 제4장에서는 이렇게 표현된 메타 데이터의 활용 방안을 분석한

다. 마지막으로 제5장에서는 본 연구의 의의와 한계를 결론으로 제시한다.

2. 통합 데이터 품질 평가 모형

데이터 품질은, 일상적인 제품의 품질과 비교하여, 무형적 특성을 측정하여야 한다는 점에서 품질에 대한 기준(또는 차원: dimension)의 설정과 실제 측정(metric)이 쉽지 않았다. 전통적으로 품질이란 사용자의 요구를 얼마나 잘 충족시키는 가로 판단된다. 즉, 서비스 품질은 사용자들이 기대하는 품질 수준과 인지된 품질 수준과의 차이로 인식되어 왔다. 여기서 품질수준을 구성하는 품질 차원에는 내구성 등과 같은 공학적인 차원도 포함되나 사용 편이성 등과 같은 주관적인 차원들도 포함한다. 이와 같이, 품질은 주관적인 차원들에 대한 정성적인 평가를 포함한다.

데이터에 대한 품질 평가 또한 데이터 품질 자체에 대한 객관적인 평가로부터 사용자들에 의한 주관적인 평가를 포함하도록 확장되어 왔다. 또한 최근에는 이러한 품질에 영향을 미치는 품질의 생성 원인을 추적하여 이를 같이 평가하는 방향으로 확대되어 왔다.

2.1 데이터에 대한 평가

데이터에 대한 평가 척도로서 가장 일차적인 관심사는 데이터가 실제를 얼마나 잘 표현하는가를 나타내는 데이터의 정확성이다. 이에 따라 데이터 품질에 대한 연구 또한 데이터의 정확성을 중심으로 주로 논의되어 왔다. 이후, 사용자의 정보 요구에 대한 정성적인 충족도를 나타내는 유용성, 완전성 등의 평가 기준이 추가되었다.

이에 따라 정보 서비스 업체 등에서 가장 일반적으로 이용하는 데이터의 품질 기준은 정확성, 유용성 및 완전성의 3가지 척도이다. 실제 현상에 대하여 사용자들이 알고자 하는 정보를 제공한다는 관점에서 데이터 품질은 무엇보다도 정확한 정보를 제공하여야 하며, 이러한 정보들이 사용자들의 용도에 유용하여야 하며, 사용자들이 찾고자 하는 모든 요소들에 대한 정보를 완전하게 제공할 수 있어야 한다는 정확성, 유용성 및 완전성의 관점에서 평가되어 왔다.

이러한 사용자 중심의 데이터 품질에 추가하여, 데이터베이스 관리자의 관점에서 데이터베이스 내부에 저장된 값들에 대한 관리 차원의 품질들이 제시되었다. 이는 사용자들에게 제시되는 값이 아니라 데이터베이스에 저장된 값 자체에 대한 평가를 주 관심으로 한다. 이는 데이터베이스에 저장된 값 자체에 대한 평가 기준으로서 대표적인 품질 기준으로는 이들 값들이 상호 서로 일치하여야 한다는 일관성, 논리적으로 서로 모순되지 않아야 한다는 무결성 등을 포함하고 있다.

이와 같이 데이터 품질은 평가 주체에 따라 사용자의 관점에서 평가하는 품질과 데이터 관리자의 관점에서 평가하는 품질의 2가지 차원에서 언급되어 왔다. 이들 평가의 주체에 따른 구분과 더불어 평가의 대상이 무엇이냐에 따라 데이터의 품질을 구분할 수 있다.

평가의 대상에 따라서는 데이터 자체에 대한 품질과 데이터를 제공하는 서비스에 대한 품질로 구분되어 왔다[2]. 여기서 데이터 자체에 대한 품질은 데이터베이스에 존재하는 데이터에 대한 평가로서 정확성, 완전성, 현행성, 일관성 등을 포함하며, 서비스에 대한 품질은 사용자들

2) 반드시 그러한 것은 아니나, 일반적으로 서비스에 대한 평가는 사용자 중심의 평가 관점과 연관성이 높으며, 데이터 자체에 대한 평가는 데이터베이스 관리자 중심의 평가 관점과 연관성이 높다고 할 수 있다.

이 느끼는 서비스에 대한 평가로서 검색성, 사용용이성, 사용자 지원성 등을 포함한다. 비슷한 구분으로 한국데이터베이스진흥센타는 데이터 품질로는 정확성, 완전성, 최신성, 포괄성, 활용성을, 서비스에 대한 품질로는 검색성, 편의성, 지원성, 시스템 성능을 제시하였다 [7][8].

이러한 품질 차원들에 대한 연구의 일환으로 Wang은 데이터 품질 차원들을 <표 1>의 4가지 카테고리로 분류하고 있다[5].

<표1> Wang의 데이터 품질 카테고리

내재적 정보 품질	Accuracy, Objectivity, Believability, Reputation
접근적 정보 품질	Access, Security
상황적 정보 품질	Relevancy, Value-added, Timeliness, Completeness, Amount of Data
표현적 정보 품질	Interpretability, Ease of understanding, Concise presentation, Consistent representation

Wang이 제시한 데이터 품질 카테고리들을 앞에서 언급한 평가 대상 (데이터 자체에 대한 품질과 데이터의 서비스에 대한 품질) 및 평가 주체 (사용자 중심과 데이터베이스 관리자 중심)에 따른 평가 기준들과 상호 비교하면 <표 2>와 같이 정리할 수 있다.

<표 2> 데이터 품질 차원의 상호 비교

내재적 정보 품질	데이터 자체에 대한 품질	사용자/데이터베이스 관리자
상황적 정보 품질		사용자
접근적 정보 품질	데이터의 서비스에 대한 품질	데이터베이스 관리자
표현적 정보 품질		사용자

2.2 데이터 품질 원인에 대한 평가

통합품질관리(TQM: Total Quality Management)는 제품 또는 서비스의 설계부터 제조 및 사후 관리까지의 전 과정을 포함한다. 따라서 완성된 제품 및 이에 대한 사용자들의 평가에만 국한하였던 품질관리를 이의 발생 원천부터 포함하도록 확장한다.

데이터의 경우에도 양질의 데이터를 서비스하기 위하여서는 최종 데이터에 대한 품질 관리뿐만이 아니라 데이터를 생성하는 데이터베이스 구조, 생성 프로세스 등의 모든 과정에 대한 평가가 같이 이루어져야 한다. 데이터에 대한 품질 관리를 이와 같이 확장할 경우, 데이터 품질은 크게 앞 절에서 언급한 데이터에 대한 평가와 더불어 데이터의 품질에 영향을 미치는 환경적 원인들에 대한 평가로 확장할 수 있다.

데이터 품질에 영향을 미치는 환경적 요인으로서 일반적으로 다음의 4가지를 제시하고 있다.

- 데이터 설계
- 데이터의 처리 프로세스
- 시스템
- 데이터 관리 정책 및 절차

이들 중에서 시스템에 대한 평가는 보안성, 접근성, 성능 등과 같은 데이터에 대한 평가와 연관하여 많이 다루어 졌다. 다만 시스템을 주 대상으로 하기보다는 이를 통하여 제공되는 서비스의 평가로 취급되었다고 볼 수 있다. 그리고 데이터 설계에 대한 평가 또한 앞 절의 데이터에 대한 평가에서 설명한 데이터베이스 관리자의 관점에서 평가되는 내재적 품질 차원들과 연관하여 많이 다루어졌다. 이와 같이, 데이터 설계와 시스템에 대한 평가는 기존의 데이터에 대한 평가 차원으로서 반영되어 왔다.

데이터의 품질에 영향을 미치는 원인으로

서 최근 많이 연구되고 있는 것이 데이터 처리 프로세스이다. 프로세스는 이전에는 응용 시스템의 일부로 분류되어 데이터 관리의 범위 밖의 문제로 취급되었었다. 그러나 최근에는 이를 데이터 관리의 수명주기에 포함시킴으로서 데이터 관리의 관점을 표현하는 처리 모형들이 제시되고 있다. Redman은 양질의 데이터 서비스를 위한 노력으로서 정보처리기능모형(Information Processing Model)을 제시하고 있으며[4], Wang은 정보제조시스템 (Information Manufacturing System) 등을 제시하고 있다[5]. 이들 모두 데이터의 처리 프로세스를 데이터 관리의 관점에서 단순화하여 표현한 모형들이다.

2.3 통합 데이터 품질 평가 기준

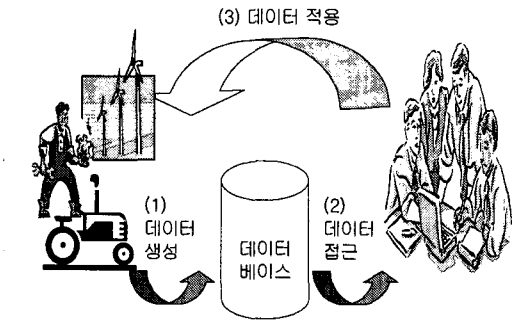
통합 품질관리의 관점에서 볼 때, 데이터의 품질 평가 또한 데이터에 대한 평가와 데이터 품질에 영향을 미치는 원인들에 대한 평가를 모두 포함함을 알 수 있다. 본 연구도 이러한 통합 품질 관리의 틀 안에서 데이터에 대한 품질 모형을 제시한다. 데이터에 대한 품질들을 종합적으로 재 정의하기 위하여 본 연구는 실제 현상의 투사 패러다임 (real-world mapping paradigm)에 기초하여 기존 연구들에서 제시한 품질 평가 기준 및 차원들을 재구성한다.

실제 현상 투사 패러다임에 기초하여 볼 때, 데이터는 실제현상을 부호로 표현한 것이며, 데이터 서비스의 목적은 실제 현상에 대한 정보를 데이터를 통하여 제공하고, 이를 이용한 사용자의 활동을 지원하는 것이라고 정의할 수 있다. 이러한 패러다임에 기초하여 데이터의 이용 주기를 살펴보면 <그림 1>과 같이 3부분으로 나눌 수 있다.

(1)데이터의 생성: 실제 현상을 나타내는 데이터를 생성하여 저장

(2)데이터의 접근: 사용자가 데이터를 제공받고 이해

(3)데이터의 적용: 데이터에 근거하여 실제 현상에 행동을 적용



(그림 1) 데이터의 이용 사이클

이들을 데이터의 이용 사이클이라고 하면, 데이터의 품질은 사이클을 구성하는 각 단계별로 다르게 평가되어야 한다. <표 3>은 데이터 이용 사이클의 각 단계별로 품질 기준과 품질 평가의 주체 또는 구성 요소들을 제시한다.

데이터 생성 단계의 품질은 데이터가 실제 현상을 얼마나 잘 나타내는가에 의하여 평가되며, 평가 시스템을 구성하는 주 요소는 데이터와 실제 현상이다. 이에 반하여 데이터 접근 단계는 데이터를 이용하고 이해하기에 얼마나 편리한가에 의하여 평가되며, 평가 시스템의 주 구성요소는 사용자와 데이터이다. 그리고 데이터 적용 단계는 사용자가 획득한 데이터가 실제 현상에 대하여 행동을 결정하고 실행함에 얼마나 유용한가를 평가하며, 이때 평가 시스템을 구성하는 주요 요소는 사용자와 실제 현상이다.

여기서 평가 시스템의 주체 또는 구성 요소는 평가의 주체 또는 평가 대상을 구성하는 개

체로서 평가 척도를 구성하는 주요 결정인자들이다. 즉, 데이터 생성 단계의 주 구성 요소가 데이터와 실제 현상이라는 것은 사용자들이 느끼는 정성적인 평가 기준은 데이터 생성 단계에서의 주요 평가 기준은 아니라는 것을 의미한다. 이에 반하여 데이터 적용 단계의 주 구성 요소가 사용자와 실제 현상이라는 것은 적용 단계에서는 사용자가 느끼는 정성적인 평가 기준들이 주요한 평가 기준으로 취급되어야 함을 의미한다.

<표 3> 실제 현상 투사 패러다임에 근거한 데이터 품질 기준

데이터의 생성	데이터가 실제 현상을 얼마나 잘 나타내는가	데이터와 실제 현상	내재적 품질
데이터의 접근	데이터를 사용자가 이해하기에 얼마나 편리한가	데이터와 사용자	표현적 품질
	데이터를 사용자가 이용하기에 얼마나 편리한가		접근적 품질
데이터의 적용	실제 현상에 대한 행동을 결정하고 실행함에 얼마나 유용한가	사용자와 실제 현상	상황적 품질

이러한 구분에 따라 제시된 평가 기준은 Wang이 제시한 4가지 품질 카테고리와 거의 일치한다. 즉, 데이터 생성의 품질 차원들은 내재적 품질 카테고리에 포함될 수 있으며, 데이터 접근의 품질 차원들은 표현적 품질 카테고리와 접근적 품질 카테고리에 포함될 수 있으며, 데이터 적용의 품질 차원들은 상황적 품질 카테고리에 포함될 수 있다. 따라서 본 연구는 데이터 품질 분류와 연관하여서는 Wang의 품질 카테고리를 그대로 원용한다.

2.4 통합 데이터 품질 평가 프레임워크

통합 데이터 품질 관리는 데이터 품질에 대한 평가와 더불어 품질에 영향을 미치는 원인들에 대한 평가를 함께 포함한다. 이에 따라 품질 평가의 대상 또한 데이터로부터 이들 원인들도 포함하도록 확장된다.

품질에 영향을 미치는 원인들로서 본 연구는 데이터베이스의 구조, 데이터의 생성 프로세스, 데이터의 관리 정책과 절차, 시스템을 포함한다. 이들은 다음에 예시된 바와 같이, 여러 가지 경로로 데이터의 품질에 영향을 미친다.

- 데이터베이스 구조는 데이터의 일관성이나 무결성과 같은 내재적 품질과 이해용이성 등과 같은 표현적 품질에 영향을 미친다.
- 생성 프로세스는 데이터를 만들고 변환, 처리하는 과정으로서 데이터의 정확성과 같은 구조적 품질과 데이터의 적시성, 완전성 등의 상황적 품질에 영향을 미친다.
- 시스템은 데이터의 검색 속도, 보안 등과 같은 접근적 품질에 영향을 미친다.
- 관리 정책/절차는 포괄적이고 간접적으로 데이터의 품질에 영향을 미친다.

이와 같이, 데이터 품질의 차원을 데이터의 이용 주기와 평가의 대상에 대한 구분하여 총괄적으로 표시한 것이 <표 4>의 통합 품질 프레임워크이다.

통합품질 프레임워크는 품질 차원들을 데이터의 이용 사이클별로 분류하여 제시하며, 데이터만이 아니라 데이터베이스 스키마, 프로세스, 시스템 및 관리 절차 등도 평가 대상으로 포함한다. 따라서 데이터에 대한 사후적 현상 평가만이 아니라 품질에 영향을 미치는 원인들에 대한 평가를 포함하도록 확장한다.

<표 4> 데이터 이용 단계와 평가 대상에 따른 통합 데이터 품질 프레임워크

데이터 생성	내재적 품질	정확성	*		*		*
		무결성	*	*	*		*
		일관성	*	*			*
데이터 접근	표현적 품질	표현일관성	*	*	*		*
		해석용이성	*	*			*
	접근적 품질	접근성		*	*	*	*
		성능				*	*
		보안				*	*
데이터 적용	상황적 품질	유용성					*
		적시성			*	*	*
		완전성		*	*		*

통합 데이터 품질 모형의 장점은 데이터 이용 사이클의 각 단계와 평가 대상별로 품질 차원을 제시함으로써 품질 평가를 위한 척도를 구체적으로 정의하기가 용이하다는 점이다. 표 4의 품질모형은 데이터 이용 사이클의 각 단계별 품질 차원과 평가 대상의 2차원 표로 제시되어 있다. 따라서 데이터 이용 사이클의 단계별로 무엇을 대상으로 하여 무엇을 평가하여야 하는가를 구체적으로 식별할 수 있다.

데이터 품질 평가는 평가 대상에 따라 크게 현상 평가와 원인 평가로 구분된다. 현상 평가는 평가 대상이 데이터이며, 사용자에게 서비스되는 데이터의 품질을 평가하는 것이다. 이에 반하여 원인평가는 데이터 품질에 영향을 미치는 원인들이 평가 대상이며, 이들 원인들의 특성을 평가하는 것이다. 예를 들면 실제 저장된 값과 참값의 일치여부를 직접 측정하는 것이 정확성에 대한 현상평가이며, 도메인이 얼마나 정확하게 정의되어 데이터 값을 입력하는 과정에서 범위를 벗어나는 데이터가 입력될 수 있는 가능성을 미연에 방지하는가를 평가하면 이는 정확성에 대한 원인 평가인 것이다.

이와 같이, 통합 데이터 품질 모형은 품질 차원과 평가 대상에 대하여 매우 다양한 현상 평가 및 원인 평가 척도들을 도출할 수 있다. 대표적인 예를 들어보면 <표 5>와 같다. 이들 척도들은 한국데이타베이스진흥센타의 품질평가 모형 및 관련 연구들로부터 정리한 것이다 [6][7][8].

3. 데이터 품질 관리 메타 데이터 모형

통합 데이터 품질 모형을 통하여 살펴본 바와 같이 종합적인 품질 관리를 위하여서는 이에 영향을 미치는 원인들에 대한 평가와 관리가 이루어져야 한다. 본 연구에서는 데이터에 영향을 미치는 요인으로서 데이터베이스 스키마, 생성/변환 프로세스, 시스템 및 관리 정책/절차 4가지를 고려한다.

데이터 관리에서 품질에 영향을 미치는 요소들은 메타 데이터로 관리되어 왔다. 본 연구에서 고려하는 4가지 요인들 중에서, 데이터베이스 스키마는 전통적으로 메타 데이터로 관리되어 왔다. 즉, 메타 데이터라는 단어가 의미하

<표 5> 통합 데이터 품질 프레임워크의 평가 척도

데이터 생성	내재적 품질	정확성	-실제 사실과의 일치 -데이터의 중복 -데이터의 오자 및 탈자	-참조 무결성 -데이터 값의 유일성 보장	-생성/가공시 오류 발생 -생성/가공 과정의 자동화 -원천 데이터의 신뢰성		-생성 프로세스의정의 -프로세스의 적용(운영성) --원시 데이터의 정호가성 검증
		무결성	-비즈니스 제약 조건의 충족 -속성 정의와 값의 일치	-도메인 정의	-생성/가공시 도메인의 적용		
데이터 접근	표현적 품질	표현 일관성	-제약조건과 값의 일치 -동일 데이터의 상호 일관성 -데이블 정의와 레코드의 일치	-코드 정의	-생성/가공시 데이터 표준의 적용		
		해석용이성		-코드 분류			
	접근적 품질	안정성				-재해 관리 방안 설계 -유지 보수 방안 설계 -시스템 모니터링	-전담 관리자 지정
		성능				-동시 처리 능력 -응답 시간	-성능 관리
		보안		-접근 보안의 단위		-물리적 접근 통제	
		편의성		-검색 기능 설계			
데이터 적용	상황적 품질	유용성					-사용자 만족도의 정기적 반영
		적시성	-최신 데이터 제공		-데이터 갱신 주기		
		완전성	-데이터의 양 -데이터의 범위 -데이터 값의 누락	-속성의 누락 -코드의 완전성	-생성/가공시 누락		-DB 모형의 관리

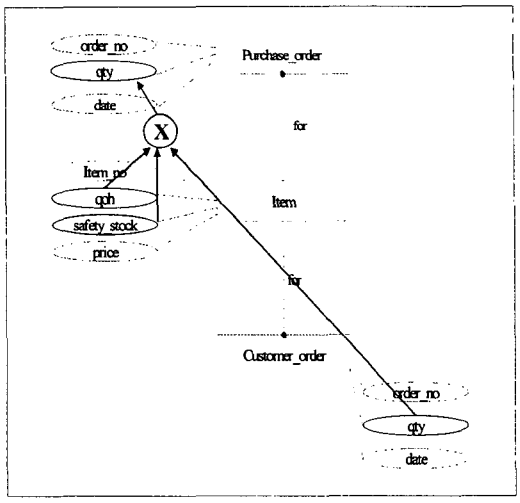
는 것처럼, 데이터에 대한 정보들 중에서 가장 일찍부터 취급되어 온 것이 데이터의 명칭, 형식, 길이 등과 같은 데이터의 저장과 관련된 데이터베이스 스키마에 대한 정보들이었다.

데이터 관리에서 품질에 영향을 미치는 요소들은 메타 데이터로 관리되어 왔다. 본 연구에서 고려하는 4가지 요인들 중에서, 데이터베이스 스키마는 전통적으로 메타 데이터로 관리되어 왔다. 즉, 메타 데이터라는 단어가 의미하는 것처럼, 데이터에 대한 정보들 중에서 가장 일찍부터 취급되어 온 것이 데이터의 명칭, 형식, 길이 등과 같은 데이터의 저장과 관련된 데이터베이스 스키마에 대한 정보들이었다.

최근에는 종합 데이터 품질 관리의 일환으로 데이터의 생성/변환 프로세스에 대한 관심이 증대하게 되었으며, 이에 대한 메타 데이터들의 관리 방안이 제기되고 있다 [4][5]. 이들 데이터의 생성/변환 프로세스에 대한 메타 데이터들은 데이터의 처리 과정을 상세히 나타내기 보다는 데이터 관리의 관점에서 처리 프로세스를 추상화 한다. 예를 들면, Redman의 정보처리기능 모형(Functions of Information processing Model)은 데이터의 처리를 Associate, Filter, Prompt, Queue, Regulate, Transit 의 6개의 기본 함수로 표현하며, Wang은 데이터의 처리과정을 data unit, vendor blocks, process blocks, quality blocks, consumer blocks로 모형화하여, vendor block으로부터 제공되는data unit이 무슨 process block 과 quality block을 거쳐 consumer block에 서비스 되는 가를 나타낸다 [4][5].

프로세스에 대한 메타 데이터들은 데이터의 처리 과정을 관리하고 평가하기 위한 메타 정보를 제공한다는 점에서 데이터 품질 관리의 영역을 크게 확장하고 있다. 그러나 이들 모형들은 데이터베이스 스키마와 독립적으로 구성됨으로써, 데이터베이스에 저장된 특정 테이블이나 필드들이 어떠한 과정을 거쳐서 생성되고 처리되는 가를 체계적으로 추적하기가 곤란하다. 이러한 프로세스에 대한 메타 데이터와 데이터베이스 스키마에 대한 메타 데이터를 결합하여 통합 제시한 메타 데이터 모형이 정보구조그래프 (information structure graph)이다 [2]

정보구조그래프는 데이터베이스 스키마를 기반으로 데이터 개체[3] 의 생성 및 변환 과정을 데이터 생성, 데이터 갱신 및 데이터 결합의 3종류의 연산자로 표시한다. 예를 들면, 재고 테이블의 현재고량(qoh)이 최초 실사 재고량에 입고량과 출고량의 차이를 가감하여 갱신된다면, (그림 2)와 같이 표시된다.

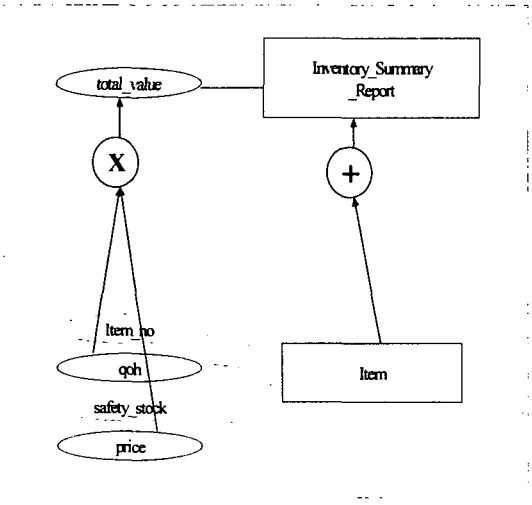


(그림 2) 정보구조그래프 (예: 현재고량(qoh) 갱신)

여기서 부호로 표시된 노드는 데이터의 생성 프로세스를 나타낸다. 프로세스 노드로부터 출력되는 화살표의 끝에 표시된 데이터 개체는 갱신되는 데이터를 나타내며, 프로세스 노드로 입력되는 화살표의 시작점에 표시된 데이터 개

3) 테이블과 속성을 포함하여 데이터 개체라고 부름

체들은 이를 갱신하기 위하여 사용되는 입력 데이터 개체들을 나타낸다. 이와 같이 정보구조그래프에서 생성되는 개체는 데이터베이스에 저장된 데이터들뿐만이 아니라 이로부터 산출되는 보고서와 같은 출력물도 포함한다((그림 3) 참조).



(그림 3) 정보구조그래프(예: Inventory Summary Report 산출)

정보구조그래프를 정형적으로 표시하면, 정보구조그래프 $G = <V, E, >$ 는 노드와 아크로 표시되는 그래프로서 다음과 같이 정의된다.

(1) V 는 노드들의 집합이며, 각 노드는 데이터 개체를 나타낸다. 그리고 E는 노드를 연결하는 아크들의 집합이다.

(2) μ(v)는 V 로부터 데이터 생성 유형의 집합 {⊗, ⓤ, ⊕} 으로의 함수로서 다음의 조건들을 만족시킨다.

(a) μ(v) = ⊙ 는 데이터개체 v 가 다른 데이터들로부터 생성되지 않은 경우, 즉 v가 그래프의 마지막 노드 (즉 leaf) 일 경우를 나타낸다.

(b) μ(v) = ⊗, ⓤ, ⊕ 는 각각 v의 생성 유형이 데이터 생성, 데이터 갱신 및 데이터 결합을 나타낸다. 이 경우 데이터 개체 v는 이를 생성하기 위하여 사용된 자식 개체들을 가진다.

정보구조그래프에서 μ(v)는 v의 데이터 생성 유형을 나타낸다. 그러나 표현의 편이를 위하여, 데이터 생성 유형을 입력 데이터개체들과 같이 표현한다. 즉,

(1) μ(v) = ⊙는 v가 기본 데이터이며, 다른 데이터 개체들로부터 생성되지 않았음을 나타낸다.

(2) μ(v) = (⊗, v_{i1}, ..., v_{in})는 v의 생성 유형이 데이터 생성 (⊗)이며 이의 입력 데이터 개체들이 v_{i1}, ..., v_{in} 임을 나타낸다.

(3) μ(v) = (ⓤ, v_{i1}, ..., v_{in})는 v의 생성 유형이 데이터 갱신 (ⓤ)이며 이의 입력 데이터 개체들이 v_{i1}, ..., v_{in} 임을 나타낸다.

(4) μ(v) = (⊕, v_{i1}, ..., v_{in})는 v의 생성 유형이 데이터 결합 (⊕)이며 이의 입력 데이터 개체들이 v_{i1}, ..., v_{in} 임을 나타낸다.

이와 같이 정보구조그래프는 데이터의 생성 구조를 같이 표현한다. 이러한 정보구조그래프를 이용함으로써 데이터 품질에 영향을 미치는 원인들인 데이터베이스의 스키마와 생성 프로세스에 대한 품질을 함께 표현할 수 있다.

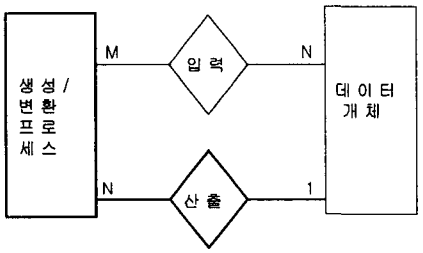
대부분의 관계형 데이터베이스에서 데이터베이스 스키마는 <표 6>과 같이 자료 사전 테이블로 표현된다.

<표 7> 자료사전 테이블

TBL (table_name, creator, table_type, ...)
COL (table_name, col_no, col_name, col_type, width, nulls, ...)

이들 각 테이블 또는 컬럼에 대하여 이의 생성구조를 정보구조그래프로 표시할 수 있다.

정보구조그래프에서 노드는 테이블 또는 컬럼과 같은 데이터 개체를 나타내며, 아크는 이들 노드를 연결한다. 이러한 정보구조그래프를 관계형 테이블로 표시하기 위하여 본 연구는 데이터 개체를 나타내는 노드와 생성 프로세스를 나타내는 노드를 구분하여 (그림 4)와 같이 모델링한다.



(그림 4) 정보구조그래프의 개체관계도

(그림 4)에 표시된 바와 같이, 산출되는 데이터 개체와 이의 생성 프로세스는 일-대-일의 관계를 가진다[6]. 따라서 이들은 하나의 노드로 표시할 수 있다. 그러나 생성 프로세스와 데이터 개체들은 이들에 대한 메타 데이터를 분리하여 기록하는 것이 보다 바람직하며, 따라서 본 연구에서는 이들을 서로 분리하여 다른 종류의 노드로 표현하였다.

(그림 4)의 데이터 모형에 기초하여 정보구조그래프는 <표 7>의 관계형 테이블들로 표시된다. 이들 각각은 데이터 개체를 나타내는 노드, 생성 프로세스를 나타내는 노드 및 입력 데이터 개체와 생성 프로세스를 연결하는 아크를 나타낸다[7].

<표 7> 정보구조그래프 테이블

ISG_data_node (data_object, data_object_type[4], ...)
ISG_process_node (output_data_object, process_name, data_creation_type, error_rate, time_lag, ...)[5]
ISG_input_arc (input_data_object, output_data_object, seq_no, ...)

이와 같이, 정보구조그래프와 데이터베이스 스키마를 이용함으로써, 데이터 품질에 영향을 미치는 원인들 중의 하나인 생성 프로세스를 데이터베이스 스키마와 동일한 메타 데이터로 형태로 표현할 수 있다.

예를 들면, (그림 2)의 품목 테이블 중에서 현재고량을 나타내는 qoh 컬럼에 대한 메타 데이터는 <표 8>의 자료사전테이블과 <표 9>의 정보구조그래프 테이블과 같이 표현된다.

4) Data_object_type 은 해당 data_object가 테이블인지 또는 속성인지를 구분한다.

5) 프로세스를 나타내는 ISG_process_node 테이블은 다양한 품질 정보들을 메타 데이터로 포함한다. 표 7에서는 대표적으로 error_rate, time_lag 만을 포함하도록 설계하였다.

6) 데이터 개체의 생성이 여러 단계의 생성 프로세스로 이루어질 수 있다. 그러나 본 연구에서는 모델링의 편의를 위하여 하나의 데이터 개체에 하나의 생성 프로세스만이 대응하는 것으로 가정한다. 그리고 만약 여러 프로세스가 이루어지는 경우에는 프로세스별로 중간 단계의 데이터 개체가 생성되는 것을 가정한다.

7) 생성 프로세스 노드는 데이터 개체와 일-대-일의 관계를 가지는 의존 개체로 모델링 하였다. 즉, 프로세스 노드는 이에 의하여 산출되는 데이터 개체가 존재하여야 한다. 따라서 이의 식별 또한 별도의 식별자를 이용하지 않고 산출되는 데이터 개체의 식별자로 정의하였다.

<표 8> 자료 사전 테이블

TBL

table_name	creator	table_type
item	Lee	table
inbound_delivery	Lee	table
outbound_delivery	Lee	table
customer_order	Lee	table
purchase_order	Lee	table
...		

COL

table_name	col_no	col_name	col_type	width	nulls
item	2	qoh	number	4	yes
item	3	safety_stock	number	4	yes
...					
inbound_delivery	2	qty	number	4	yes
...					
outbound_delivery	2	qty	number	4	yes
...					
customer_order	2	qty	number	4	yes
...					
purchase_order	2	qty	number	4	yes
...					

생성 프로세스에 대한 메타 데이터를 데이터베이스 스키마와 동일한 형식으로 표현함으로써 시스템, 관리 정책/절차 등도 생성 프로세스와 연관하여 표현할 수 있다. 즉, 데이터의 품질에 영향을 미치는 요인으로서 시스템이나 관리 정책/절차 등의 역할은 결국 프로세스를 통하여 품질에 반영된다.

<표 9> 정보구조그래프 테이블 (예시)

ISG_data_node

data_object	data_object_type
item	table
item.qoh	column
item.qoh_measured	column
inbound_delivery	table
inbound_delivery.qty	column
outbound_delivery	table
outbound_delivery.qty	column
purchase_order	table
purchase_order.qty	column
customer_order	table
customer_order.qty	column
...	...

ISG_arc

input_data_object	output_data_object	seq_no
inbound_delivery.qty	item.qoh	1
outbound_delivery.qty	item.qoh	2
item.qoh	purchase_order.qty	1
item.safety_stock	purchase_order.qty	2
customer_order.qty	purchase_order.qty	3
...	...	...

ISG_process_node

output_data_object	process_name	data_creation_type	error_rate	time_lag
inbound_delivery.qty	inbound count	-	0.005	0
outbound_delivery.qty	outbound count	-	0.005	0
item.qoh_measured	qoh manual count	-	0.02	1 day
item.qoh	qoh update	데이터 갱신	0.0001	1 hour
item.safety_stock	safety stock level entry	-	0.005	1 day
purchase_order.qty	order qty calc.	데이터 생성	0.0001	0
customer_order.qty	CO entry	-	0.01	0
...	...	...	...	...

따라서 이들 원인들에 대한 메타 데이터들 또한 <표 10>과 같이 정보구조그래프의 프로세스 테이블을 참조로 표현될 수 있다.

<표 10> 정보구조그래프를 포함하는 통합 메타 데이터 테이블들

system (system_name, ...)
system_ISG_process(system_name, output_data_object, process_name, error_rate, time_lag, ...)
policy (policy_name, ...)
policy_ISG_process(policy_name, output_data_object, process_name, error_rate, time_lag, ...)

IV. 통합 품질 관리 메타데이터의 활용

통합 데이터 품질 관리 모형은 데이터베이스 스키마에 대한 메타 데이터뿐만이 아니라 프로세스에 대한 메타 데이터를 같이 포함한다. 따라서 이는 통합 데이터 품질 관리를 위한 종합 정보를 제공하며, 데이터 관리 정책 및 절차를 실행하는 기준이 된다.

통합 품질 관리 메타 데이터의 활용 분야를 예시하면, 데이터의 정확성이나 적시성과 같은 품질 척도의 관리 및 이에 영향을 미치는 요인의 파악들을 거론할 수 있다.

4.1 품질 척도의 관리

정보구조그래프 테이블들은 데이터베이스 스키마와 같이 동일한 관계형 데이터베이스 테이블로 자료사전 데이터베이스에 저장될 수 있다. 따라서 이를 이용할 경우, 데이터의 생성 과정 (즉 프로세스)에서 발생하는 여러 종류의 품질 척도들을 효과적으로 기록하고 검색, 활용할 수 있다.

예를 들면, 현 재고량의 갱신 프로세스에서 오류가 (그림 5)와 같이 발생한다고 가정하자. 이 경우, 이들 오류 발생률은 이미 앞에서 제시되었던 <표 9>의 정보구조그래프 테이블들에 저장된다.

4.2 품질 발생 원인의 파악

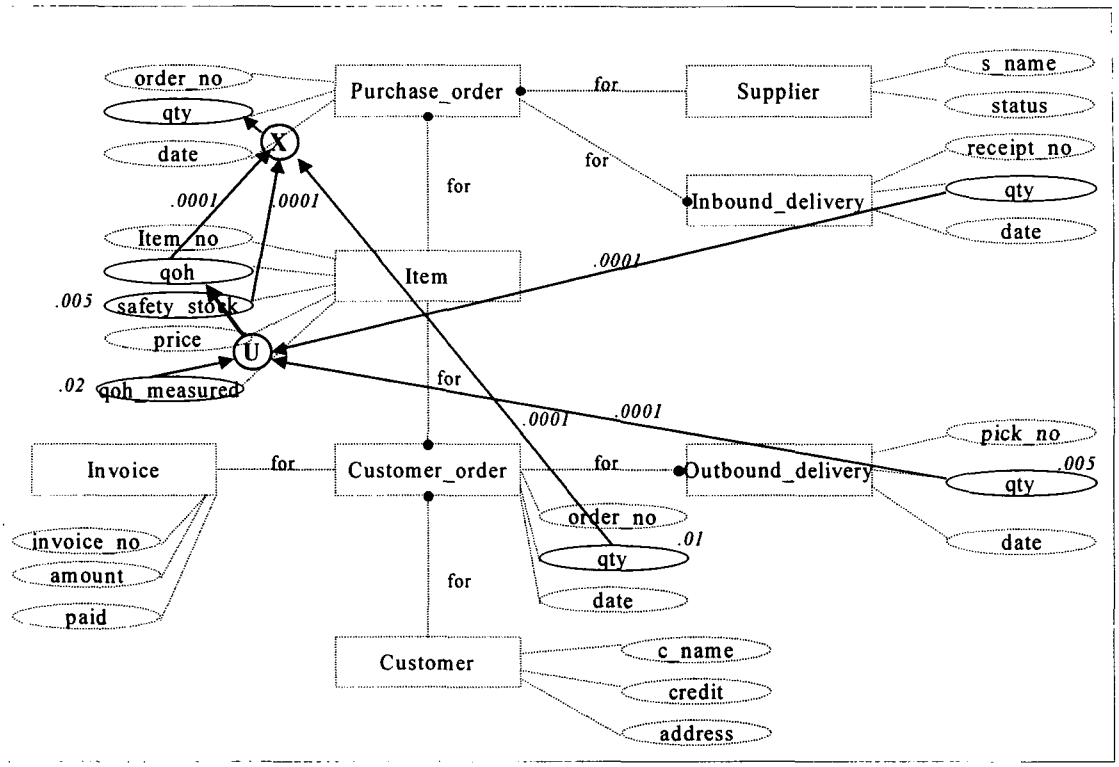
<표 8>과 <표 9>등에 저장된 데이터베이스 스키마 및 프로세스에 대한 메타 데이터들로부터 item 테이블의 qoh 컬럼의 오류 발생률에 영향을 미치는 요인들은 다음과 같이 검색된다.

```
SELECT output_data_object,
       input_data_object, error_rate
FROM   ISG_arc
CONNECT BY
PRIOR  input_data_object =
       output_data_object
START WITH input_data_object =
          item.qoh'
```

위의 질의는 item.qoh를 생성하기 위하여 사용되는 모든 입력 데이터 개체들과 이들 개체의 오류 발생률을 검색하며, 검색 결과는 <표 11>과 같다. 이로부터 오류 발생의 주 원인이 item 테이블의 qoh_measured 컬럼임을 알 수 있다. 이와 같이, 정보구조그래프 테이블에 저장된 메타 데이터들로부터 품질 관리에 필요한 정보들을 검색할 수 있으며, 이로부터 품질에 영향을 미치는 주요 원인을 식별할 수 있다.

품질에 영향을 미치는 주요 원인의 파악과 더불어 이를 이용하여 역으로 이들 원인들로부터 생성되는 모든 데이터 개체들을 검색할 수 있다.

```
SELECT input_data_object,
       output_data_object
FROM   ISG_arc
CONNECT BY
PRIOR  output_data_object =
       input_data_object
START WITH output_data_object =
          'item.qoh_measured'
```



(그림 5) 오류 발생률을 포함한 정보구조그래프

검색 결과는 <표 12>와 같이 나타내며, 이를 통하여 item 테이블의 qoh_measured 컬럼으로부터 item 테이블의 qoh 컬럼과 purchase_order 테이블의 qty 컬럼이 생성됨을 추적할 수 있으며, 따라서 이들 또한 높은 오류 발생률을 가짐을 예측할 수 있다.

<표 12> 데이터 품질 메타 데이터의 검색 결과

input data object	output data object	output of output data object
item.qoh_measured	item.qoh	
item.qoh_measured	item.qoh	purchase_order.qty

V. 결론

지식 정보를 축적하기 위하여서는 양질의 데이터를 장기적으로 축적 관리할 수 있는 품질 관리 체계가 구축되어야 한다. 이러한 품질 관리 체계의 구축에서 가장 기반을 이루는 것이 체계적인 품질 관리 모형이며, 이를 위한 메타 데이터의 관리이다.

전통적으로 데이터 품질 관리는 사용자에게 제공되는 데이터에 대한 품질들이 주로 다루어져 왔다. 그러나 통합 품질 관리의 관점에서 데이터 이용 사이클의 전 과정에 걸쳐 품질에 영향을 주는 모든 원인들에 대한 관리가 중요하게 되었으며, 이를 위한 연구들이 최근 데이터 품질 관리의 중심을 이루고 있다. 이러한 관점

에서, 본 연구는 데이터 품질 관리를 위한 통합 프레임워크를 제시하고 이에 대한 메타 데이터 표현 방법을 제시하였다.

본 연구는 통합 품질 관리를 위한 메타 데이터들을 표현하기 위하여 정보구조그래프를 활용한다. 정보구조그래프는 데이터베이스를 구성하는 테이블과 속성들 사이의 생성 프로세스를 표현하는 그래프로서 이를 이용할 경우, 데이터베이스 스키마에 대한 메타 데이터와 생성 프로세스에 대한 메타 데이터를 동일한 형태로 표현할 수 있다. 또한 데이터 품질에 영향을 미치는 시스템이나 관리 정책/절차 등과 같은 원인들에 대한 품질 정보 또한 동일한 형식으로 표현할 수 있다.

이와 같이 데이터 품질에 영향을 미치는 여러 원인들에 대한 메타 데이터들이 동일한 형식으로 표현되고 관리됨으로써, 데이터 품질에 대한 통합 관리가 실제로 가능함을 예시적으로 살펴보았다. 나아가 메타 데이터들을 이용하여, 데이터 품질에 영향을 미치는 이들 원인들에 대한 검색과 관리 방법 또한 살펴보았다.

본 연구는 데이터 품질에 대한 메타 데이터를 통합적으로 표현하기 위한 실험적인 연구이다. 따라서 이의 확장 가능성과 적용 범위들에 대한 추가적인 연구가 필요하다. 이러한 추가적 연구를 위하여서는 품질 원인들에 대한 평가 항목들을 정형화하고 이들을 표현하기 위한 메타 데이터의 형식들이 연구되어야 할 것이며, 이를 활용한 사례 연구들이 수반되어야 할 것으로 생각된다.

참 고 문 헌

[1] Dolk, D. R. and Kirsh II, R.A., "A Relational Information Resource Dictionary System", *Communications of the ACM*, Volume 30, Number 1, January 1987, 48-61

[2] Lee, C.Y., A Knowledge Management Scheme for Meta-Data: An Information Structure Graph, Proceedings of PACIS 2001, June 20-22, Seoul, Korea, 933-947.

[3] Redman, T.C., "Improve Data Quality for Competitive Advantage", *Sloan Management Review*, 1995 Winter, 99-107.

[4] Redman, T.C., *Data Quality for the Information Age*, Artech House, 1996.

[5] Wang, R.Y., "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, Vol. 41 No. 2, Feb. 1998, 58-65.

[6] 김창화, 정보자원사전에 대한 서술논리 표현과 관리, 정보기술과 데이터베이스 저널, 제5권 1호, 1998년 9월, 13-37.

[7] DPC, DB 품질평가모델 연구보고서, 2002. 12.

[8] DPC, DB 품질평가모델 확장 개발 연구보고서, 2003. 7.

■ 저자소개

이 춘 열

저자는 서울대학교 산업공학과 학사, University of Michigan에서 경영정보학박사 (Computer & Information Systems 전공)를 수여받았으며, 이후 한국통신 연구개발단을 거쳐 현재 국민대학교 비즈니스 IT전문대학원에 재직중이다.

◈ 이 논문은 2003년 8월 25일 접수하여 1차 수정을 거쳐 2003년 10월 21일 게재확정되었습니다.