데이터 품질관리의 이해

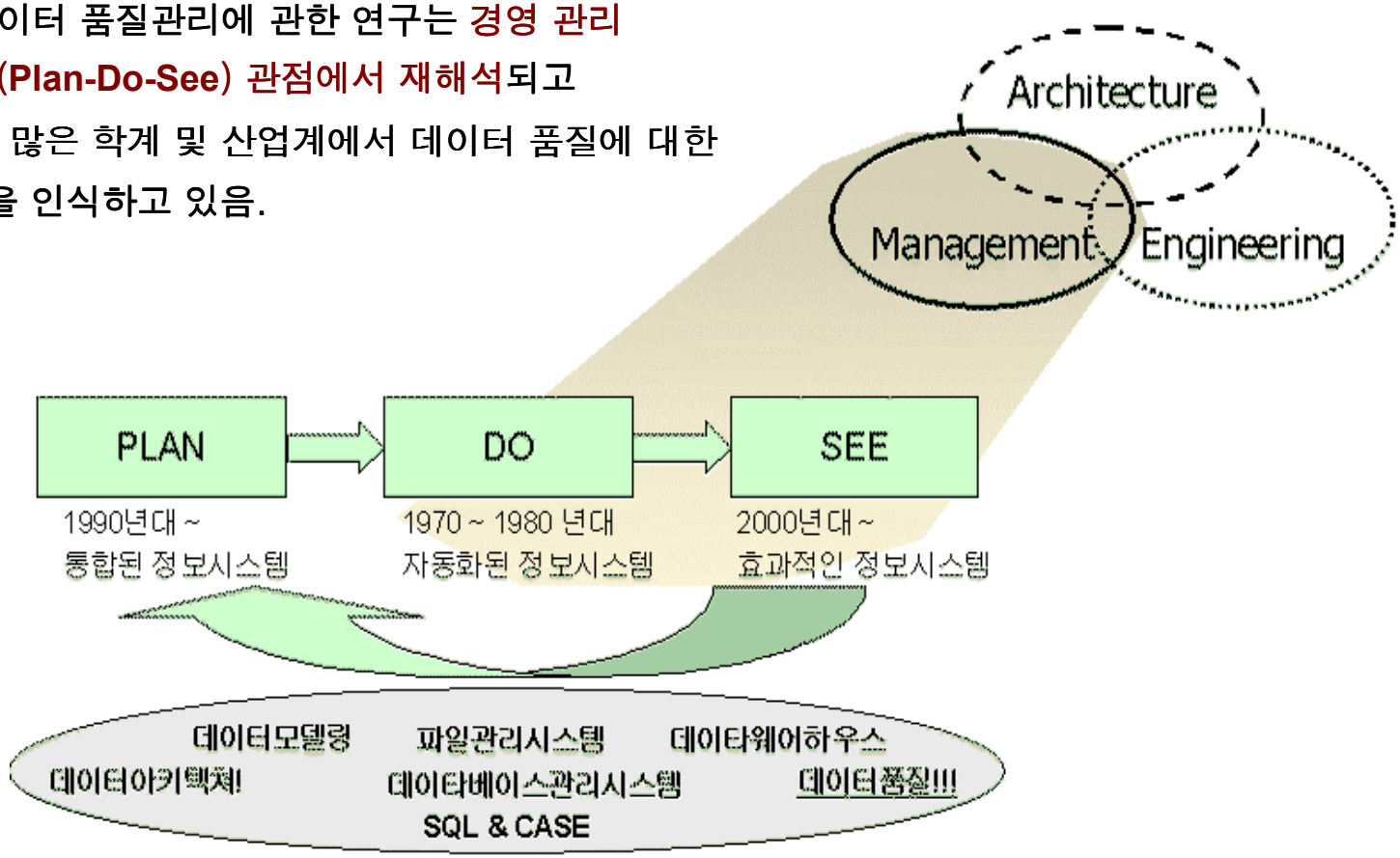
박주석
경희대학교

목차

Ⅰ. 데이터 품질 정의
Ⅱ. 데이터 품질관리 필요성 및 사례
Ⅲ. 데이터 품질관리 프레임워크
Ⅳ. 데이터 품질관리 교육 요약
참고1. 국내외 연구동향
참고2. 데이터 품질관리 성숙도 모형

Ⅰ. 데이터 품질 정의 – 데이터 품질관리의 역할

- 데이터에 관한 연구는 지난 30년 동안 IT 분야에 이론적으로나 산업적으로 엄청난 변화를 일으켜 옴.
- 현재 데이터 품질관리에 관한 연구는 **경영 관리 사이클 (Plan-Do-See) 관점에서 재해석**되고 있으며, 많은 학계 및 산업계에서 데이터 품질에 대한 중요성을 인식하고 있음.



Ⅰ. 데이터 품질 정의 – 데이터 품질이란 무엇인가

“ Consistently

meeting

all knowledge worker and end-customer

expectations ”

through data and data services

to accomplish enterprise and customer objectives.

- Larry P. English -

- Consistently : 데이터 품질은 일부 지식 노동자가 아닌 전체 지식 노동자의 기대에 부응하는 것으로 데이터 간의 일치성 유지가 중요함
- Meeting : 모든 데이터는 완전무결할 필요는 없으며, 사용자 Needs에 부합하는 수준이어야 함
- Knowledge worker & End customer : 데이터를 자신의 업무 수행 시 얼마나 도움이 되는 지를 근거로 데이터 품질이 만족스러운 지 아닌 지를 결정하는 Information customer를 의미함
- Expectation : 데이터 품질의 실제 목표는 사용자와 이해관계자 만족을 증대시키는 것

Ⅱ. 데이터 품질관리 필요성 및 사례 – 저품질 데이터의 위험(1)

- **8년 동안 2,300만 달러를 더 지불한 미국 시민** – 미국 연금 계산 오류로 인하여 Nashvill 시민 일부는 8년 동안 2,300만 달러를 초과하여 지불하였으며, 반면에 다른 사람들은 같은 기간 동안에 2,600만 달러를 적게 낸 것으로 밝혀졌다. (The Tennessean, 1998년 03월 21일 기사 일부)

- **세금 고지서 수취인 불분명으로 반환** – 1992년, 미국 국세청은 96,000 장의 세금 고지서가 수취인이 거주하지 않는 다는 이유로 반송되었다고 밝혔다. 이는 거주지 데이터의 오류로 인한 문제로 데이터 회계 감사에서 밝혀졌다.

- **금융계좌 주민번호 398만개 잘못 기재… 이사소득에 대한 금융과세 못해** – 금융회사 고객 원장의 주민등록번호가 잘못 기재된 것은 금융기관과 감독기관의 관리 및 감독 소홀이 주된 이유... 특히 동일인이 2개의 주민번호로 금융거래를 하거나, 하나의 번호를 동시에 2~3명이 사용해 온 것이어서, 선의의 피해를 부르거나 금융사고나 범죄에 악용 됐을 가능성이 높다(조선일보 2004년1월7일 기사 일부)

- **호적정보시스템 주민등록번호 10명중 1명 꼴 엉터리** – 호적정보시스템에 기록된 정보 가운데 971만여 명의 주민등록 번호가 누락되거나 잘못 기재되는 등 오류가 있는 것으로 파악… 호적상 주민번호에 오류가 있을 경우 국민들은 대출이나 구직 또는 혼인신고 등에 필요한 호적등본을 제때 발급 받지 못 하는 불편을 겪게 돼.(한국일보 2004년7월8일 기사 일부)

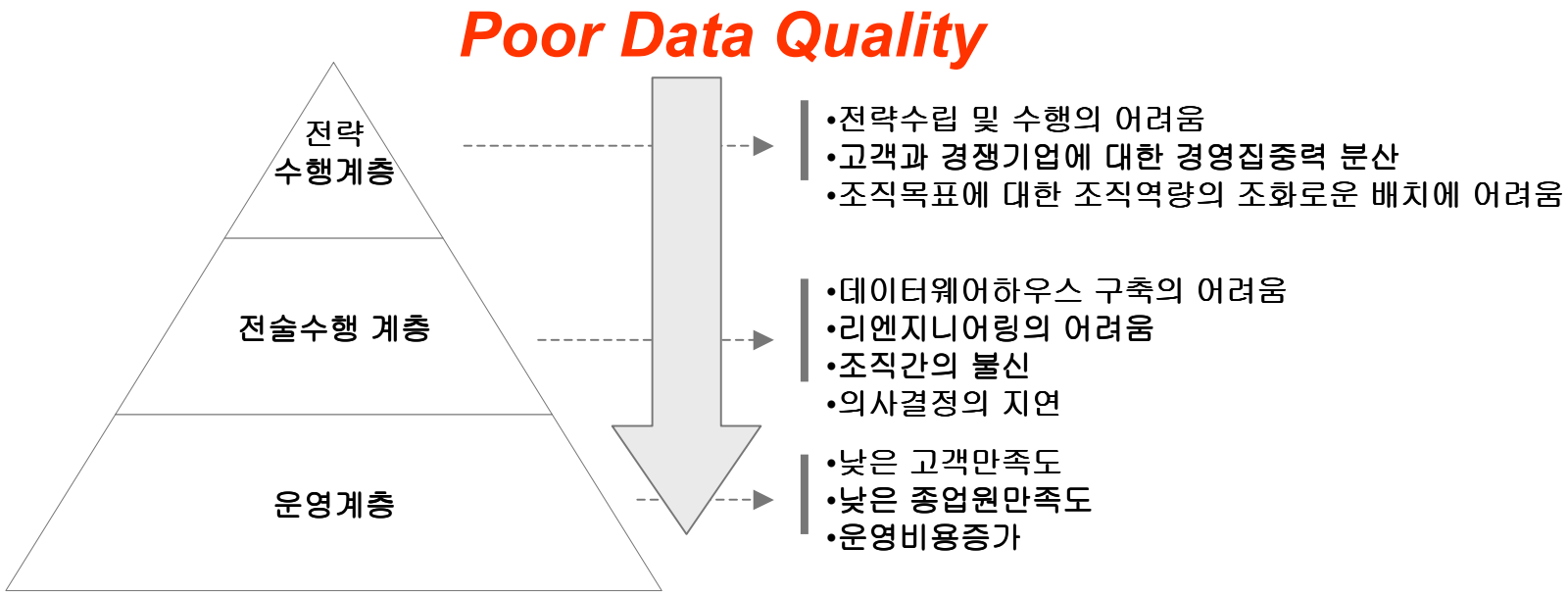
Ⅱ. 데이터 품질관리 필요성 및 사례 – 저품질 데이터의 위험(2)

- **국내 공공 기관 DB 관리 허술** – 공공기관의 데이터베이스 관리체계가 허술한 것으로 나타났다. 이는 최근 전자정부 프로젝트의 일환으로, 공공기관 간 데이터 공동 활용이나 통합을 위한 시스템 구축사업에도 상당한 걸림돌로 작용할 것으로 전망돼 대책마련이 시급한 것으로 지적됐다. 한국데이터베이스진흥센터가 지난 6~9월까지 46개 공공기관을 대상으로 실시한 '2004 데이터베이스 품질진단결과' 데이터와 이에 대한 관리 수준은 100점 만점에서 58.7점으로 나타났다. 부문별로는 데이터 영역의 품질 수준은 평균 60.6점으로 상당수 공공기관의 데이터베이스에 엉뚱한 데이터가 들어 있거나 당연히 있어야 할 데이터가 없거나, 동일한 내용에 대해 내부 시스템 별로 서로 다른 데이터를 갖고 있는 것으로 나타났다.

  또 관리 프로세스 영역은 평균 56.1점으로 대다수 공공기관이 아직 데이터 표준이나 데이터 모델을 체계적으로 관리하기 위한 프로세스를 갖추지 못한 것으로 조사됐다. 이 밖에 데이터베이스 관련 시스템 분야는 평균 59.4점으로 나타났다. 김선영 선임연구원은 "일반적으로 민간기업의 잘못된 데이터로 인해 발생되는 손실은 해당 기업이 떠안겠지만 공공기관의 데이터 품질로 인한 피해는 고스란히 행정서비스의 수혜자인 국민에게 돌아가 그 심각성이 크다"고 말했다. 실제로 최근 호적정보시스템에서 10명 중 1명꼴로 데이터 오류가 발견되고 금융사의 금융계좌 100개 가운데 1개꼴로 잘못된 주민등록번호가 사용되는 사례는 데이터베이스 관리의 허점을 적나라하게 보여준 대표적인 사례라고 진흥센터 측은 설명했다. 또 이 같은 국내 상황은 미국은 연방정부가 제공하는 정보의 품질을 보장하기 위해 '데이터 품질법(Data Quality Act)'을 제정하고 예산관리국이 실행 가이드라인을 만들어 제공하는 것과 대조적이라고 진흥센터는 덧붙였다. (전자신문 2004년 10월 12일 기사)

Ⅱ. 데이터 품질관리 필요성 및 사례 – 저품질 데이터의 위험(3)

- 데이터 품질 저하로 인해 발생된 데이터 손실 비용, 재 작업 비용 등이 기관이나 기업의 연간 예산이나 수익의 10~25%에 이름(**DATA QUALITY** 2001년 9월호)
- 데이터 품질 저하로 인한 고객 불만, 제품 관련 소송, 재 작업 등의 비용이 평균 매출액 25~30%에 이르는 것으로 분석됨(품질전문가 Joseph M.Juran)



■ 데이터 품질 법(Data Quality Act) 제정 과정

PRA (Paperwork Reduction Act)

- 80년 대, 미 연방 정부에서 각 주(State), 지방 정부, 중소 사업장, 개인 등에 관한 데이터 수요가 증가하자, 요청 데이터를 통제하기 위한 목적으로 제정

OIRA (Office of Information and Regulatory Affairs)

- 미 예산국(OMB, Office Management and Budget)이 각 공공기관들의 Paperwork 감독자들을 관리하기 위한 목적으로 설립

DQA (Data Quality Act)

- 2000년, 데이터에 관한 환경 변화를 감안하여, 기존 PRA를 개정하여 새로이 제정.
- 2002년, 최종 데이터 품질 지침(Information Quality Guidelines)을 제정하여 각 미 연방 정부의 산하 기관 및 배포 대상 기관들에게 배포 및 시행

■ 데이터 품질 법(Data Quality Act) 목적

정부 기관 웹 사이트를 통하여 제공되는 데이터의 신뢰성과 정확성 수준을 높이기 위한 데이터 품질 관리에 관한 표준화된 지침을 제공하며,

표준 지침에 따라 공공 목적으로 배포되는 일체의 데이터는 데이터의 신뢰성과 정확성 일정 수준을 유지해야 함.

■ 데이터 품질 법(**Data Quality Act**) 적용 대상

PRA를 적용 받는 일체의 기관

행정부, 국방부, 공기업 및 정부 출연에 의한 설립 기관, 감시 및 감사 등의 업무를 수행하는 기관 등이 포함.

■ 데이터 품질 법(Data Quality Act) 상의 데이터 품질

Utility (이용 편이성)

- 제공 데이터는 데이터 이용자(단체 및 개인 등)들의 이용 만족도가 높아야 한다.
- 제공 데이터는 데이터 이용자들이 쉽게 구할 수 있어야 한다.

Integrity (안전성 및 신뢰성)

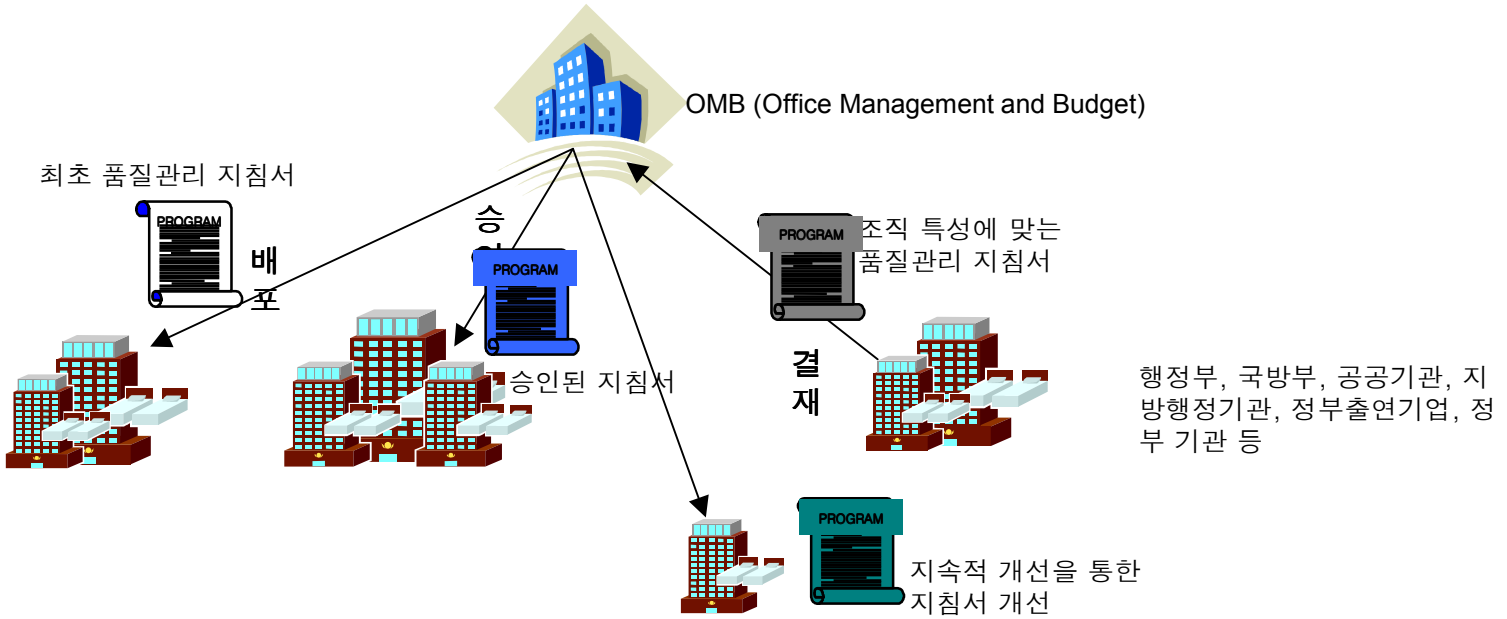
- 데이터에 대한 사용자 접근 권한 관리를 통한 비 권한자들로부터 데이터를 보호해야 한다.
- 기술적(Technical), 관리적(Managed), 물리적(Physical) 보안 강화를 통하여 데이터를 보호해야 한다.

Objectivity (공익성)

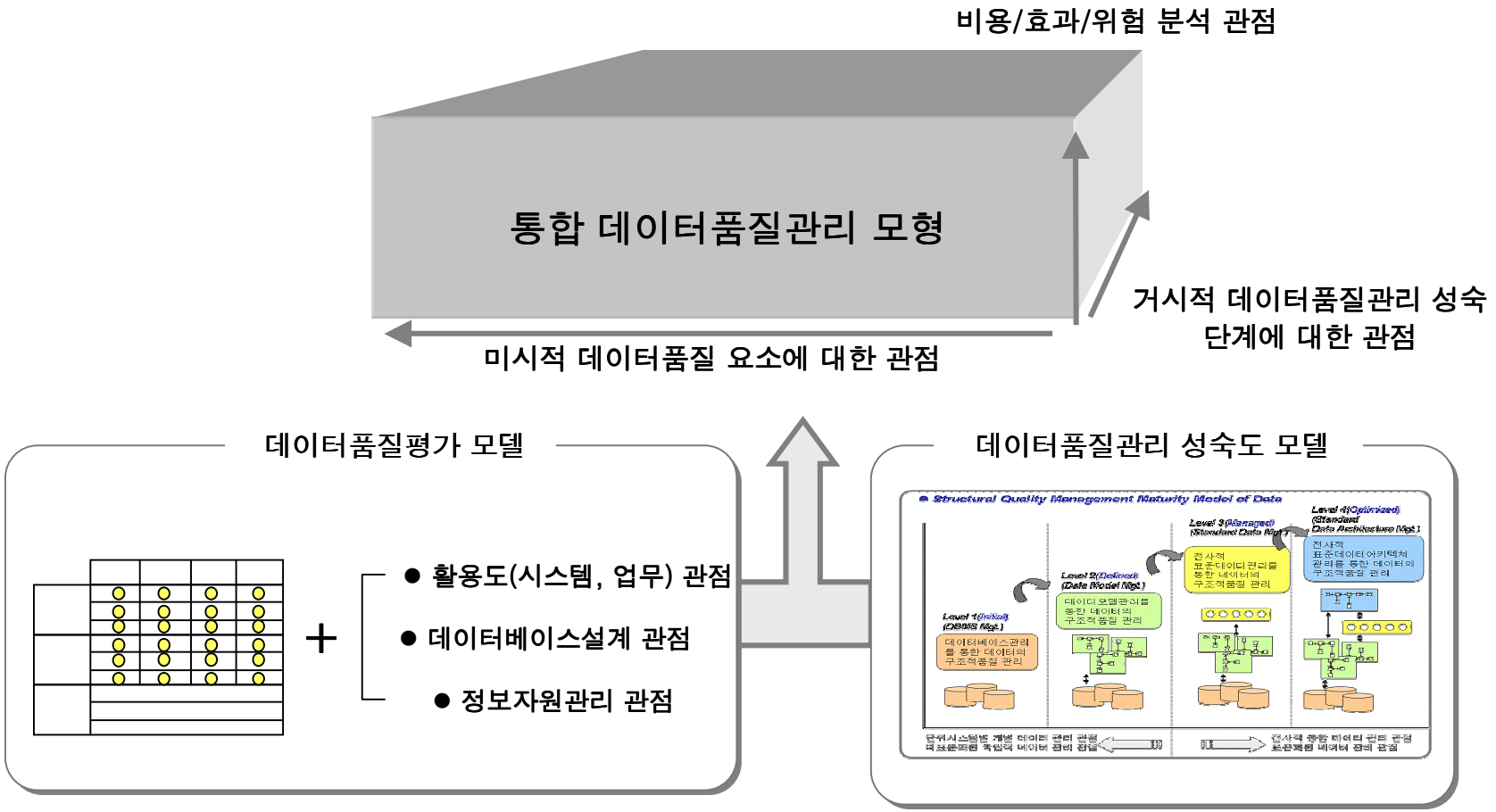
- 데이터는 정확하고, 분명하며, 완전해야 하며, 편향되지 않아야 한다.
- 데이터는 근거 및 출처가 분명해야 하며, 근거 및 출처는 과학적 혹은 통계적 기법에 의하여 증명되어야 한다.

■ 미 예산국 데이터 품질관리 지침

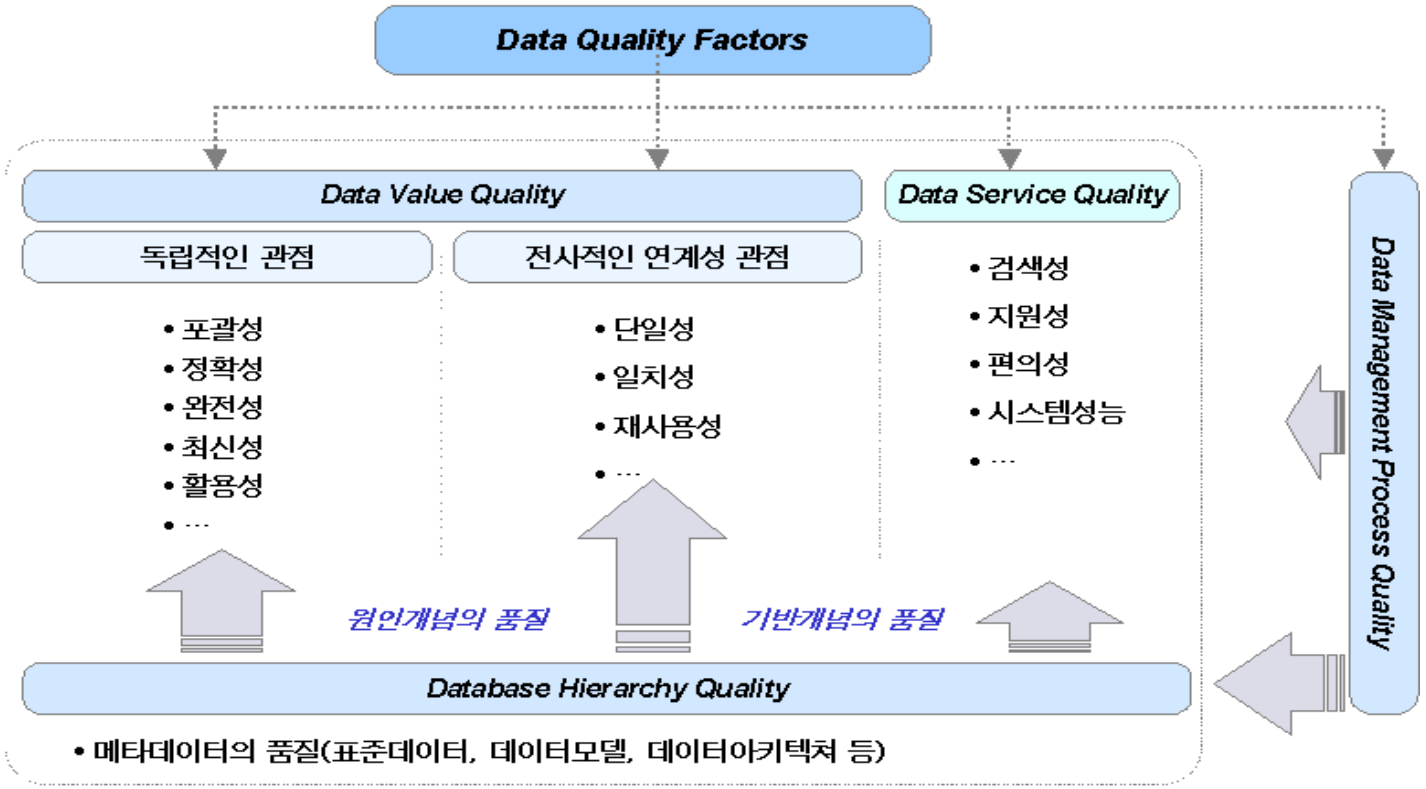
- 미 예산국(**OMB**)에서 각 공공기관에 배포한 데이터 품질관리 지침을 바탕으로, 각 공공기관은 각자의 조직 특성에 맞는 데이터 품질관리 지침을 제작.
- 기관별로 제작된 지침은 미 예산국에서 최종 승인.
- 각 공공기관은 데이터 품질관리 지침을 바탕으로 품질관리 활동을 수행하며, 또한 지속적으로 품질관리 활동을 수행하면서 개선사항이 발견되는 대로 이를 반영하여, 탄력적이고 현실적인 데이터 품질관리 지침을 유지하도록 함.



■ 프레임워크 개발 전략


비용/효과/위험 분석 관점
통합 데이터품질관리 모형
거시적 데이터품질관리 성숙 단계에 대한 관점
미시적 데이터품질 요소에 대한 관점
데이터품질평가 모델
● 활용도(시스템, 업무) 관점
● 데이터베이스설계 관점
● 정보자원관리 관점
데이터품질관리 성숙도 모델
Structural Quality Management Maturity Model of Data

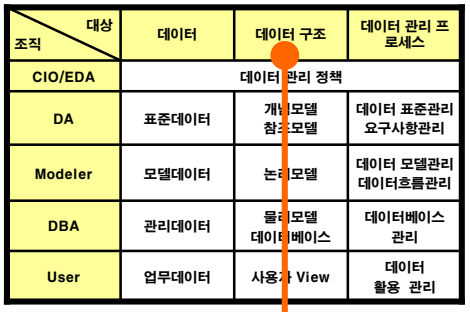
■ 데이터 품질 요소에는 크게 데이터 값(**Data Value**), 데이터 서비스(**Service**), 데이터 구조(**Data Hierarchy**), 데이터 관리 프로세스(**Data Management Process**) 등이 있으며, 이러한 요소들은 서로 연계되어 조직 데이터 품질에 영향을 주고 있어, 통합적이고 체계적인 품질 관리 노력을 필요로 함.



Ⅲ. 데이터 품질관리 프레임워크 – 프레임워크(버전 1.0)

대 상 / 조 직	데이터	데이터구조	데이터관리프로세스
CIO/EDA (개괄적 관점)	데이터 관리 정책		
DA (개념적 관점)	표준 데이터	개념모델 참조모델	데이터 표준 관리 요구사항 관리
Modeler (논리적 관점)	모델 데이터	논리모델	데이터 모델 관리 데이터 흐름 관리
DBA (물리적 관점)	관리 데이터	물리모델 데이터베이스	데이터베이스 관리
User (운용적 관점)	업무 데이터	사용자 View	데이터 활용 관리

■ 대상 정의



대상 / 조직	데이터	데이터 구조	데이터 관리 프로세스
CIO/EDA	데이터 관리 정책		
DA	표준데이터	개념모델 참조모델	데이터 표준관리 요구사항관리
Modeler	모델데이터	논리모델	데이터 모델관리 데이터흐름관리
DBA	관리데이터	물리모델 데이터베이스	데이터베이스 관리
User	업무데이터	사용자 View	데이터 활용 관리

품질관리 대상	정의
데이터	기관 및 기업의 비전이나 목표를 달성하기 해서 사용되어지는 전산화된 데이터 및 전산화에 필요한 데이터
데이터 구조	데이터가 담겨져 있는 모양/틀로서, 데이터를 취급하는 관점(view)에 따라 구조가 달라지는데, 데이터 구조는 사용자 View(양식, 보고서, 화면, 정표 등), 모델(개괄,개념,참조,논리,물리), 데이터베이스 파일의 형태로 보여진다.
데이터 관리 프로세스	데이터 및 데이터 구조의 품질을 안정적으로 유지 개선하기 위한 활동으로 절차, 조직, 인력 등을 포함

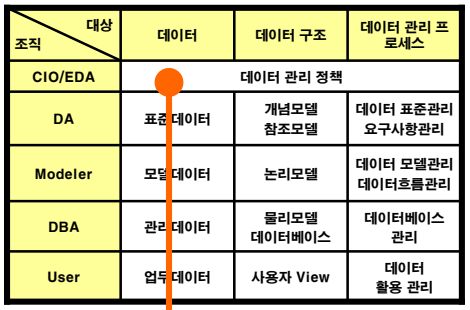
Ⅲ. 데이터 품질관리 프레임워크 – Y축: 품질관리 조직

■ 조직 정의

조직 \ 대상	데이터	데이터 구조	데이터 관리 프로세스
CIO/EDA	데이터 관리 정책		
DA	표준데이터	개념모델 참조모델	데이터 표준관리 요구사항관리
Modeler	모델데이터	논리모델	데이터 모델관리 데이터흐름관리
DBA	관리데이터	물리모델 데이터베이스	데이터베이스 관리
User	업무데이터	사용자 View	데이터 활용 관리

조직	정의
CIO/EDA (개괄적 관점)	데이터관리 총괄 데이터관리 정책 및 지침 마련 데이터관리자간 이슈사항 조정
DA (개념적 관점)	표준 개발 및 형상관리, 검증/ 표준화 절차 수립, 운영 전사데이터 모델 통합 데이터요구사항에 대한 정리 및 FDA 지원
Modeler (논리적 관점)	해당기능영역의 데이터요구사항 및 이슈사항 조정과 통합 해당기능영역의 비즈니스 요건을 토대로 데이터 모델링 수행 표준 확인 및 적용
DBA (물리적 관점)	데이터베이스 디자인 데이터베이스와 데이터의 형상관리 수행 데이터베이스의 모니터링 및 튜닝, 보안관리
User (운용적 관점)	서비스되는 데이터 및 운영/분석데이터에 대한 활용 데이터에 대한 추가요건 요청

■ 데이터



조직 \ 대상	데이터	데이터 구조	데이터 관리 프로세스
CIO/EDA	데이터 관리 정책		
DA	표준데이터	개념모델 참조모델	데이터 표준관리 요구사항관리
Modeler	모델데이터	논리모델	데이터 모델관리 데이터흐름관리
DBA	관리데이터	물리모델 데이터베이스	데이터베이스 관리
User	업무데이터	사용자 View	데이터 활용 관리

조직	대상	
	데이터	정의
DA 개념적 관점	표준데이터	정보시스템에서 사용되어지는 용어, 명칭, 도메인, 코드, 데이터 모델에 대한 표준을 지정하여 사용한다. 이러한 표준 관련 데이터를 말한다.
Modeler 논리적 관점	모델데이터	데이터 모델을 관리,운용하기 위해서 필요로 되어지는 데이터를 말한다. 세부적으로는 데이터 모델 데이터, 데이터 베이스 구조 데이터를 말한다.
DBA 물리적 관점	관리데이터	데이터를 효과적으로 관리, 유지하기 위한 프로세스에서 파생되어지는 데이터이다. 그 예로서는 데이터 흐름 데이터, 정합성 검증 데이터, 백업 내역 데이터, 데이터 사용 데이터 등이 있다.
User 운용적 관점	업무데이터	Business를 영위하기 위해 필요로 되어지는 업무데이터를 말한다. 예를 들면 계약데이터, 생산데이터, 상품데이터,고객데이터 등이 있다.

■ 데이터 구조

조직 \ 대상	데이터	데이터 구조	데이터 관리 프로세스
CIO/EDA	데이터 관리 정책		
DA	표준데이터	개념모델 참조모델	데이터 표준관리 요구사항관리
Modeler	모델데이터	논리모델	데이터 모델관리 데이터흐름관리
DBA	관리데이터	물리모델 데이터베이스	데이터베이스 관리
User	업무데이터	사용자 View	데이터 활용 관리

조직	대상	
	데이터 구조	정의
DA 개념적 관점	개념모델	업무 요건을 충족하기 위해서 데이터의 주제영역과 핵심 데이터 집합을 정의하고 관계를 정의한 모델
	참조모델	업무영역별, 주제 영역별 표준 데이터 집합, 관리 항목들이 표기되어 재 사용이 가능한 데이터모델
Modeler 논리적 관점	논리모델	개념 모델을 상세화 하여 논리적인 데이터 집합, 관리항목, 관계를 정의한 모델
DBA 물리적 관점	물리모델	논리모델을 DBMS의 특성 및 성능을 고려하여 구체화시킨 모델.
	데이터베이스	물리 모델을 적용하여 구축된 실제 데이터가 저장되어지는 데이터 저장소
User 운용적 관점	사용자 View	사용자가 사용하는 데이터 활용 화면 및 출력물

■ 데이터 관리 프로세스

조직 \ 대상	데이터	데이터 구조	데이터 관리 프로세스
CIO/EDA	데이터 관리 정책		
DA	표준데이터	개념모델 참조모델	데이터 표준관리 요구사항관리
Modeler	모델데이터	논리모델	데이터 모델관리 데이터 흐름관리
DBA	관리데이터	물리모델 데이터베이스	데이터베이스 관리
User	업무데이터	사용자 View	데이터 활용 관리

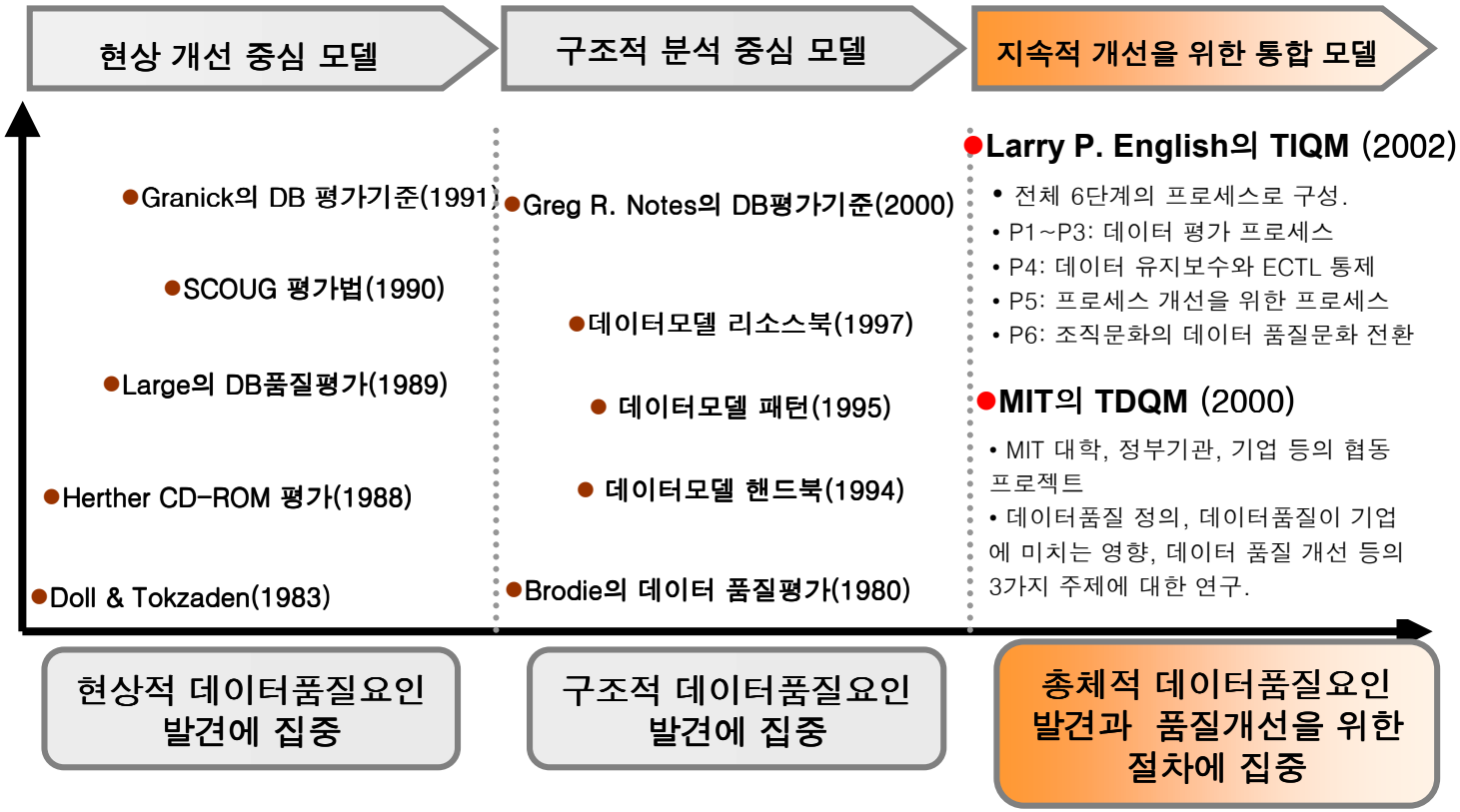
조직	대상	
	관리 프로세스	정의
CIO 개괄적 관점	데이터 관리정책	기관 및 기업의 비전과 목표를 달성하기 위해 필요한 데이터 확보 계획과 확보된 데이터를 효과적으로 관리,유지하기 위한 체계 및 계획
DA 개념적 관점	데이터 표준관리	데이터 및 데이터 관련 요소의 명칭, 정의, 도메인, 코드(코드 분류기준 및 코드값 포함) 표준을 정의하고 관리하는 작업을 말함
	요구사항관리	데이터와 관련된 사용자의 요구를 수집하고 분류하여 데이터 및 데이터 구조, 프로세스에 반영하는 작업을 말함.
Modeler 논리적 관점	데이터 모델관리	데이터 모델링을 통해 설계된 모델을 표준 및 DB 구조에 맞게 유지, 관리하는 작업으로 모델 형상관리를 포함함.
	데이터 흐름관리	원천데이터(문서, Text, DB 등)를 수기로 생성하거나 추출, 변환, 적재, 가공을 통해 목표 데이터베이스에 저장하는 작업으로 모니터링 작업을 포함함
DBA 물리적 관점	데이터베이스 관리	원활한 데이터 서비스를 위해 안정적으로 데이터베이스를 유지하는 작업으로 백업, 보안, 튜닝, 재구성, 모니터링 등의 작업을 포함함.
User 운용적 관점	데이터 활용 관리	데이터 활용 수준과 사용자 만족도를 높여주기 위한 작업.

Ⅳ. 데이터 품질관리 교육 요약

일자	시간	교육주제	교육내용
1일째	09:30~10:30 (60분)	Overview	- 데이터 품질관리의 이해
	10:40~12:00 (80분)	데이터 관리정책 수립 및 운용 : CIO 관점의 품질관리	- 데이터 관리정책 및 지침 수립 - 데이터 관리정책 운용
	13:00~16:00 (180분)	데이터 표준관리 및 전사적 모델 관리 : DA 관점의 품질관리	- 전사 데이터 모델 관리 - 데이터 표준 및 표준화 절차 관리 - 데이터 요구사항 관리
	16:00~17:30 (90분)	데이터 활용 관리 : User 관점의 품질관리	- 업무데이터 활용 관리 - 운영・분석데이터 활용 관리
2일째	09:30~12:30 (180분)	데이터 모델 관리 : Modeler 관점의 품질관리	- 데이터 모델 관리 - 데이터 흐름 관리
	13:30~17:00 (210분)	데이터베이스 관리 : DBA 관점의 품질관리	- 데이터베이스 설계 - 데이터베이스와 데이터 형상관리 - 모니터링, 튜닝 및 보안관리

참고1. 국내외 연구 동향 – 해외

■ 초기의 데이터베이스 품질 평가 모델들은 현상 분석 중심에서 점차적으로 구조적 분석 중심 모델로 바뀌었으며, 최근에는 **TDQM, TIQM** 등과 같이 지속적 개선을 위한 통합모델로 연구가 진행되고 있습니다.

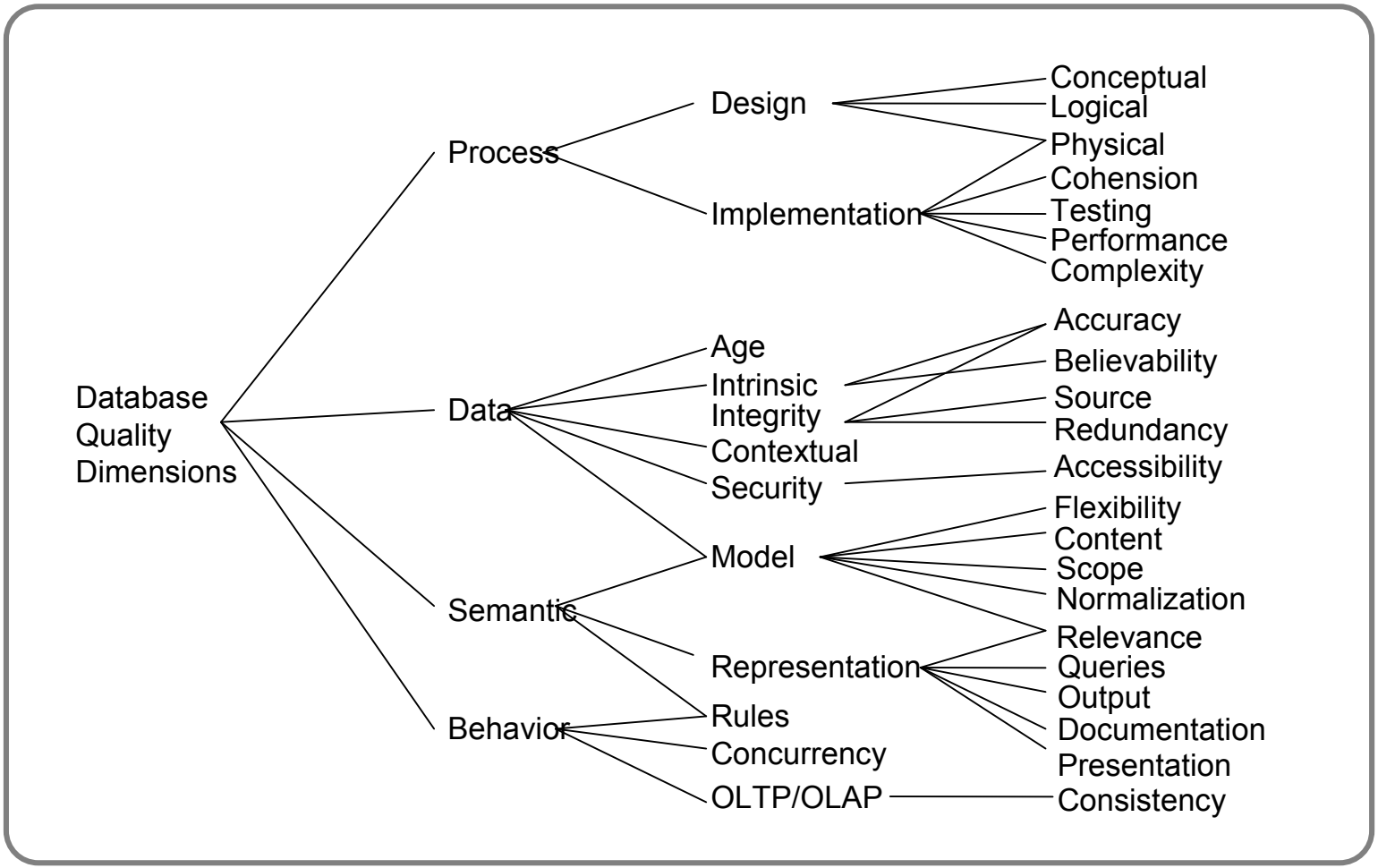


자료원: DPC 2003, 개정: 경희대 박주석 2004

참고1. 국내외 연구 동향 – 해외

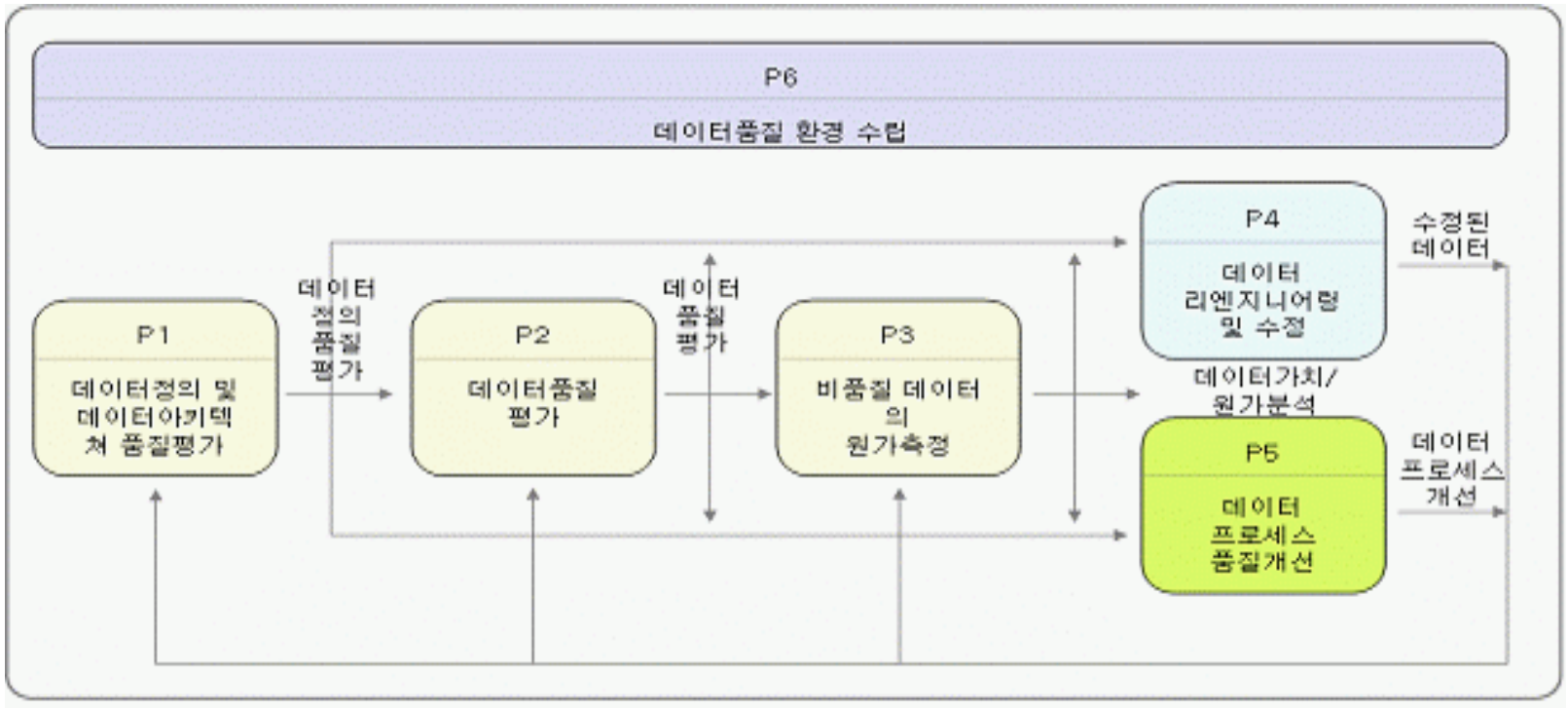
	평가기준	평가관점	평가지표
데이터값의 품질	포괄성	DB에 수록되어 있는 데이터가 양적으로 충분한지에 대한 평가를 목적으로 하며, 수록 레코드의 수의 범위 측면에서 포괄성을 측정한다	– 수록레코드의 규모 – 수록 데이터의 범위
	정확성	데이터의 오류가 없는 정도를 의미하며, 데이터값과 표현 측면에서의 정확성을 측정한다.	– 데이터값의 정확성 – 레코드의 중복성 – 디렉토리 속성의 정확성 – 데이터 표현의 일관성
	완전성	DB를 구성하는 레코드의 완성도를 의미하며, 데이터 구조와 값에 있어서의 완전성을 측정한다.	– 레코드구성필드의 완전성 – 데이터필드 값의 완전성
	최신성	갱신을 통해 DB가 최신의 데이터를 유지하는 정도를 의미한다.	– 업데이트 주기 – 최신자료의 수
	활용성	주이용자 계층의 성격에 비추어볼때, DB데이터가 이들의 목적 달성을 위해 효용가치를 제공할 수 있는 정도를 평가 한다.	– 이용자 관심과의 부합도 – 이용자 효용가치
데이터서비스의 품질	검색성	DB가 제공하는 검색방법의 다양성과 편의성 및 결과값의 적절성을 평가한다.	– 검색방법의 다양성 – 검색결과의 적절성 – 검색결과의 출력성
	지원성	DB이용안내 및 이용자지원을 위한 커뮤니케이션 등이 원활히 이루어지고 있는가를 평가한다.	– 이용안내의 충실성 – 커뮤니케이션의 원활성
	편의성	사용자에게 요구하는 과업 및 화면의 기본적인 구성 등 DB 시스템 인터페이스 측면을 평가한다.	– 사용자과업의 용이성 – 인터페이스의 용이성
	시스템성능	시스템 속도 및 안정성 등을 평가한다.	– 시스템 속도의 적절성 – 시스템 안정성

John A. Hoxmeie의 데이터품질 영역



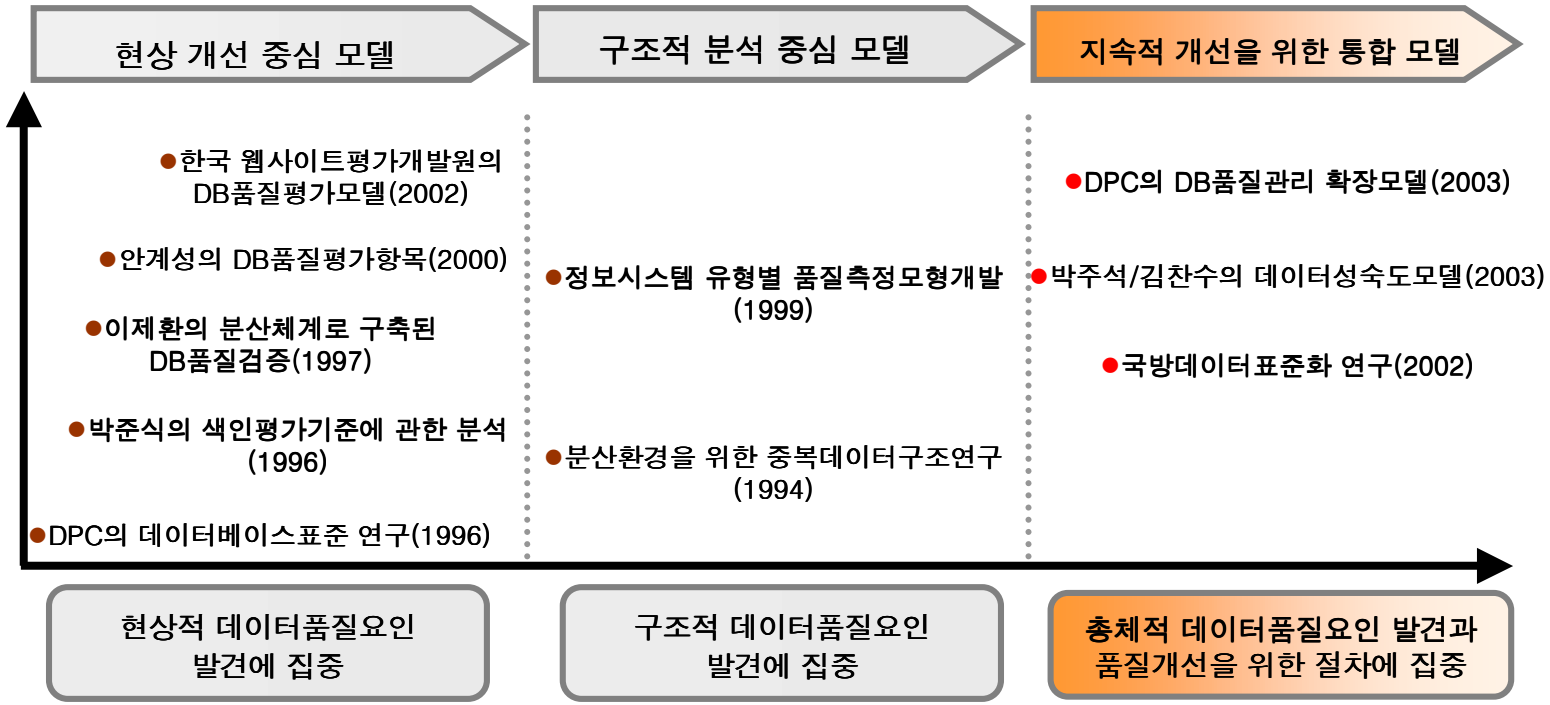
참고1. 국내외 연구 동향 – 해외

- **Larry P. English의 TIQM (Total Information Quality Management)**



참고1. 국내외 연구 동향 – 국내

- 국내에서는 90년대 말부터 데이터품질의 중요성이 부각되면서 이에 대한 연구가 진행되어 왔습니다. 특히 많은 국내 기업들이 데이터웨어하우스를 구축하고 CRM을 구현하면서 국내에서도 지속적인 개선을 위한 데이터품질에 관한 연구가 활발한 것으로 예상됩니다.



참고1. 국내외 연구 동향 – 국내

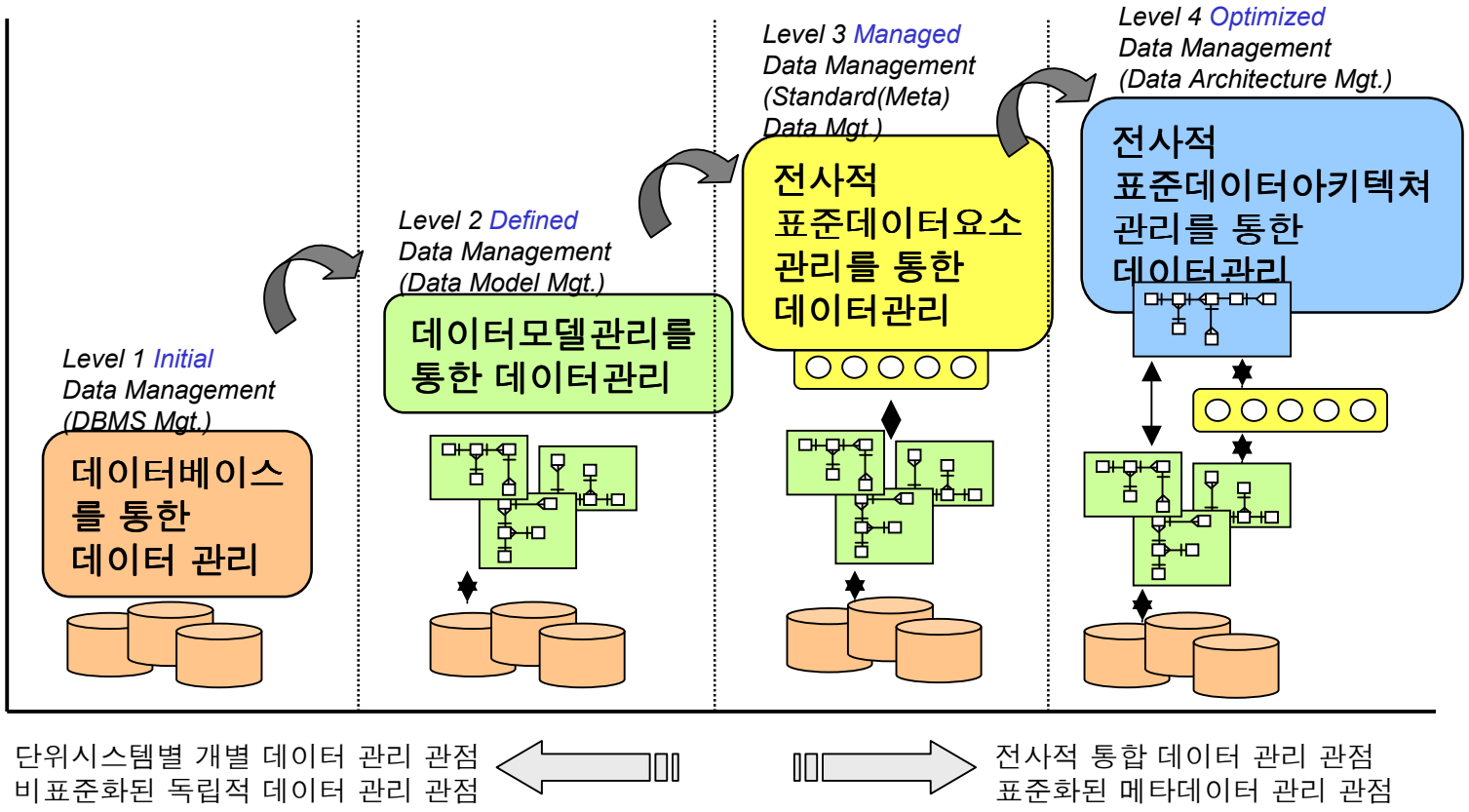
■ 투이컨설팅이 개발한 DPC의 데이터품질관리모형은 데이터품질관리의 모든 요소를 고려한 프레임워크를 도출했다는 점에서 의의가 있으나 실용성 측면에서는 개선 필요성이 있음

대 상		품질관리 요소			
대분류	중분류	전략	운영	설계	구축
데이터	데이터 값	● 데이터 관리 전략	● 데이터 요건	● 데이터 논리특성	● 데이터 물리 특성 ● 데이터
	데이터 구조		● 데이터표준 ● 도메인정의 ● 개념모델	●논리 모델	● 물리모델 ●데이터베이스
	데이터 흐름		● 데이터 생성/가공 요건	● 데이터 생성/가공 체계	● 데이터 생성/가공 시스템
데이터관리 프로세스		● 데이터관리 정책 수립	● 데이터 요구 사항관리 ● 데이터 표준 관리	● 데이터 모델 관리	● 데이터 베이스 관리 ● 데이터 생성/가공관리
시스템		하드웨어			
		소프트웨어			
		네트워크			

자료원 : DPC 2003

참고2. 데이터 품질관리 성숙도 모형

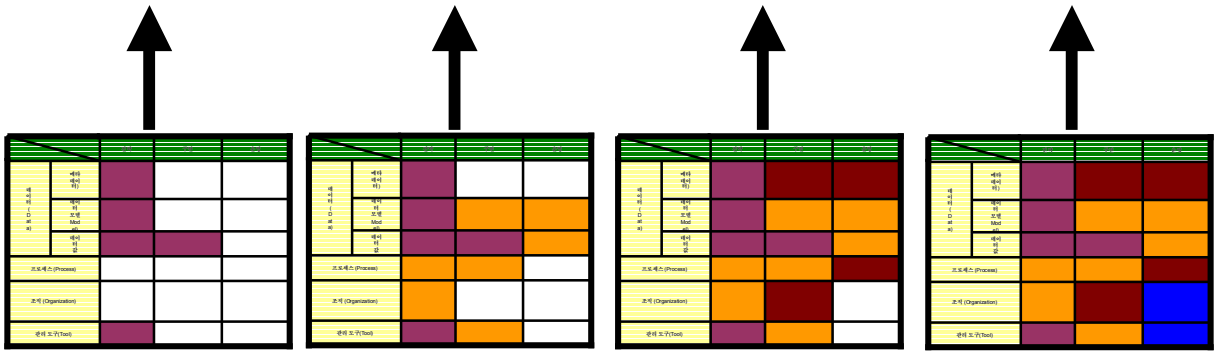
DMM Model (Data Management Maturity Model)



참고2. 데이터 품질관리 성숙도 모형 – 레벨 0부터 4까지

- 데이터 품질관리 프레임워크 내의 품질관리 대상을 기본으로 품질관리 활동을 수행한다. 수행할 때, 각 자의 조직 규모, 체계, 특성 등을 파악하여 조직이 어느 성숙도 단계에 해당하지는 파악한다.

레벨 0	레벨 1	레벨 2	레벨 3	레벨 4
애플리케이션에 의한데이터 관리	**DBMS에 의한 데이터 관리**	**모델링에 의한 데이터 관리**	**전사적 표준에 의한 데이터 관리**	**데이터 아키텍처에 의한 데이터 관리**
정의 : 애플리케이션 프로그램에 의해 파일을 정의하고 관리하는 단계	정의 : 데이터베이스시스템을 통하여 데이타베이스 품질을 관리하는 단계	정의 : 데이터베이스 구축 방법론을 사용하여 논리적인 데이터모델과 물리적인 데이터모델을 관리하는 단계	정의 : 통합의 관점에서 데이터 표준화를 통한 전사적인 데이터 품질을 관리하는 단계	정의 : 전사적인 데이터 아키텍처 관리를 통한 각 데이터 계층을 유기적으로 관리하는 단계



Q & A

Contact Info : jspark@khu.ac.kr