행정정보 데이터세트 기록 이관 시 데이터 보정 및 품질 개선 방법 연구
- 데이터웨어하우스 ETT 경험을 기반으로

임진희* · 조은희**

* 한국국가기록연구원 학술처장, 명지대학교 기록정보과학 전문대학원 겸임교수
** 명지대학교 기록정보과학 전문대학원 박사과정.

www.kci.go.kr

[국문초록]

공공 부문의 정보시스템 의존도가 점차 높아지면서 행정정보시스템에 축적되는 데이터세트 기록의 관리와 활용에 관한 다양한 방안이 모색되고 있다. 행정정보 데이터세트를 아카이브 시스템이나 공유서버로 이관할 때 데이터 보정이나 품질 개선의 요구가 발생할 수 있다. 이 논문의 목적은 데이터웨어하우스 구축을 위해 데이터를 추출하여 변형 후 전송하는 절차와 방법을 참조하여 이관하는 행정정보 데이터세트 기록의 보정 및 품질 개선 방법을 제시하는 것이다.

이 논문에서는 데이터세트 기록 이관 시 검토할 필요가 있는 전형적인 데이터 보정 및 품질 개선 사례로 (1)추출 시 데이터세트 수량과 유효값 확인, (2)일관된 코드값의 부여를 위한 코드 변환, (3)복합정보의 컴포넌트화, (4)날짜데이터의 정밀도 결정, (5)데이터 표준화, (6)코드값의 설명정보 (7)메타데이터 확보 등 7가지를 제시하고 각각의 처리방법을 제안하고 있다.

데이터세트 기록 이관 시 적용하는 데이터 보정 및 품질 개선 기준은 데이터세트를 생산하는 행정정보시스템의 데이터 품질요건으로 활용할 수 있다.

주제어 : 데이터세트, 전자기록 이관, 테이터 보정, 품질 개선, 데이터웨어하우스, EIT

www.kci.go.kr

1. 머리말

1.1 연구 목적

전자정부의 추진으로 한국 정부기관들은 상당히 높은 수준으로 업무의 디지털화를 이루었다.[1] 대부분의 정부기관들은 주요 공문서 처리를 위해 전자문서시스템을 사용하고 있으며[2], 기관의 업무 홍보 및 시민과의 소통을 위해 홈페이지를 운영하고 있다.[3] 또한 각종 민원처리와 내부 업무처리를 효율적으로 하기 위해 다양한 행정정보시스템을 구축하여 운영하고 있으며[4], 업

1) 전자문서 유통기관 현황(2007 전자정부사업연차보고서, p.62)

구분	대상	2002	2003	2004	2005	2006	2008.1
중앙행정기관	59	46	46	55	57	58	59
지방자치단체	248	-	240	252	250	248	248
공공기관	410	-	-	-	-	73	210
기타	-	-	-	-	24	271	292
총계	-	46	286	307	550	660	809

행정기관간 전자결재율 및 전자문서 유통률(2007 전자정부 사업 연차보고서)

구분	2002	2003	2004	2005	2007
전자결재율(%)	89.5	92.6	96.3	97.7	98.4
전자문서 유통건수(백만건)	-	-	21	31	42
전자문서 유통률(%)	78.1	85.9	95.7	95.8	97.8

2) 신전자문서시스템은 2004년 1월부터 중앙행정기관을 시작으로 사용하기 시작 (전자신문, 정부기관 신전자문서시스템 보급사업 추진, 2003년 8월 16일자) http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=105&oid=030&aid=0000040154

3) 전자정부법의 제2장 전자정부서비스의 제공 및 활용 부분에서 전자적 민원 처리, 구비서류의 전자적 확인, 민원처리, 행정정보의 전자적 제공 등에 사항을 규정하고 있다.

4) 전자정부사업백서(행정자치부, 2003-2007년)에 따르면 분야별로 도입된 주요 행정정보시스템은 다음과 같다.

www.kci.go.kr

무관리를 위해 온나라시스템을 도입하고 있다.[5] 이처럼 공적 업무 수행 과정에서 정보시스템에 대한 의존도가 높아지면서 업무활동에 관한 대량의 디지털 정보가 시스템에 축적되고 있다. 이 중에서 전자문서 기록의 관리체계는 일차적으로 수립되어 실행되고 있으며[6], 데이터세트 기록이나 웹기록에 관한 관리 방안이 적극적으로 모색되고 있다[7].

데이터세트 기록의 경우 국가기록원은 2007년 '행정정보시스템 데이터세트 기록관리 연구용역'[8]을 통해 데이터세트 기록의

분야별	주요 행정정보시스템의 예
재정	국가재정정보시스템, 디지털예산회계시스템, 지방재정정보시스템, 지방교육재정정보시스템, 국방재정정보시스템
전자지방정부	시도행정정보시스템(18개 공통행정업무), 정책결정지원시스템, 시군구행정정보시스템(31개 공통행정업무),
감사	e-감사시스템
형사·사법	형사사법통합정보시스템
인사	자치단체인사행정정보시스템, 전자인사관리시스템(e-사람), 상훈관리시스템
외교·통상	외교통상정보시스템
국정·행정	온라인실시간국정관리시스템, 행정심판인터넷서비스시스템
국가안전관리	긴급구조표준시스템, 국가재난관리정보시스템, 소방예방정보시스템, 재난정책지원시스템
건축·토지	인터넷건축행정정보시스템, 부동산정보관리시스템
복지	국가복지정보시스템
식·의약품	식의약종합정보서비스시스템, 농축수산물안전관리시스템
국세	국세통합시스템

5) 행정안전부에서 제공하는 온나라시스템에 대한 소개 팜플렛 http://www.mopas.go.kr/gpms/view/jsp/download/userBulletinDownload.jsp?userBtBean.bbsSeq=1010550&userBtBean.ctxCd=1002&userBtBean.orderNo=2 [2010.04.15]

6) 2007년 개정된 공공기록물관리법령은 정부기관들이 사용하는 정보시스템 중 전자문서시스템과 업무관리시스템을 기록생산시스템으로 특정하여 기록의 이관 지침을 별도로 명시하고 있다.

7) 또한, 그 밖의 행정정보시스템들에 대해서도 기록의 생산시스템이라는 지위를 부여함으로써 이메일, 웹페이지와 웹자원, 데이터세트, 모바일 메시지 등 다양한 유형의 기록을 의무적으로 관리하도록 하는 법제도적 근거를 갖추고 있다.

8) 국가기록원, 「행정정보시스템 데이터세트 기록관리 연구용역 보고서」, 2007

www.kci.go.kr

개념정립과 선진사례 분석, 데이터세트 기록의 아카이빙을 위한 절차 방법론 등을 정리하였고, 2009년에 후속사업인 '데이터세트 및 비표준문서 기록관리체계 시범구축 사업'[9]을 통해 데이터세트 아카이브의 위상에 대한 검토를 수행하고 시범 시스템을 구축하였으며, 2010년 이후에는 기관별로 행정정보시스템 데이터세트의 아카이빙을 수행할 계획을 갖고 있다. 이는 2005년에 작성된 기록관리혁신로드맵[10]의 첫 번째 과제인 '공공 업무 수행의 철저한 기록화'에서 세부과제로 제시되었던 '데이터세트의 기록화'가 제도적·실무적 집행 단계에 접어들었음을 의미한다.[11] 그러므로 이제는 행정정보 데이터세트 기록을 아카이브로 이관하여 관리하고 서비스하는 각 단계에서 데이터세트 기록의 특성상 고유하게 발생할 수 있는 구체적인 문제를 상세히 파악하고 이에 대한 대안을 찾아나가는 문제 기반(problem-based)의 연구가 필요한 시점이다. 이 논문은 데이터세트 기록과 연관한 전형적인 문제의 하나로 데이터세트 기록을 아카이브 등으로 이관하는 과정에서 데이터의 보정과 품질 개선이 필요함을 사례 중심으로 제시하고 이에 대한 대안을 제시하려는데 목적이 있다.

이 논문은 다음과 같은 조건을 전제로 한다. 먼저, 데이터세트 기록이란 기록의 내용(content)이 데이터베이스 테이블과 칼럼의 집합으로 구성된 것이다. 둘째, 데이터세트 기록은 특정 어플리케이션을 통해 생성, 변경, 삭제, 조회되는 데이터의 일부이다.

---

9) 국가기록원, 「데이터세트 및 비표준문서 기록관리체계 시범구축 사업」, 2009

10) 대통령기록관에서 아카이빙하여 제공하는 정부혁신지방분권위원회의 웹사이트 자료 참고. http://innovation.pa.go.kr/briefing/view.htm?id=1063&page=2

11) 2009년의 사업 과정에서 초반에 4차례에 걸친 워크숍을 통해 데이터세트 아카이브 구축에 관한 여러 이슈들을 제기하고 논의하였으나 정책과 지침이 확정되어 공표되는 단계까지는 이르지 못하였다.

www.kci.go.kr

셋째, 데이터세트 아카이브는 데이터베이스 기반의 저장소를 중심으로 구축되며, 데이터세트 기록을 테이블과 칼럼의 집합체로 이관하여 저장·관리한다. 넷째, 데이터세트 생산시스템의 데이터베이스관리시스템과 아카이브의 데이터베이스관리시스템은 이종일 수 있다. 이러한 전제를 기반으로 이 논문에서는 행정정보 데이터세트를 아카이브로 이관하는 과정에서 데이터의 보정과 품질 개선이 필요한 전형적 사례를 제시하고 사례별 보정 및 개선 방법을 제시하고자 한다.[12] 보정과 개선 사례와 방법은 지난 십여 년 간 기업과 공공기관 영역에서 다수 구축되었던 데이터웨어하우스의 경험을 응용하고 있다.

1.2 연구 동향

이 논문은 데이터세트 기록의 이관 과정에서 데이터 보정 및 품질 개선의 필요성과 방법 제시에 초점을 두고 있다. 먼저 국내 데이터세트 기록관리에 관한 연구동향과 데이터세트 기록을 포함하는 포괄 범주로서 전자기록의 이관에 관한 선행연구를 살펴보고자 한다.

먼저. 국내 데이터세트 기록관리에 관한 연구는 다음 두 가지 영역에서 살펴볼 수 있다. 첫 번째 영역은 개별 행정정보시스템의 데이터세트 기록화 방법에 초점을 맞춘 연구이다. 교육행정정보시스템(NEIS)[13][14], 디지털예산회계시스템[15], 조달업무시스

12) 데이터의 보정과 품질 개선 행위가 기록으로서의 무결성을 해치는 것이 아니냐는 이 논문의 직접적 논제에서 벗어난다. 다만, 이 논문에서는 디지털 정보의 무결성은 비트스트림의 자체의 불변성을 의미할 수 있으나 기록으로서 데이터세트의 무결성은 비트스트림이 나타내고자 했던 의미 혹은 메시지의 불변성을 의미하는 것임을 전제로 한다.

13) 임미숙, 「교육행정정보시스템의 기록관리 기능분석 - 학교생활기록부를 중심

www.kci.go.kr

템[16], e-사람[17] 등의 행정정보시스템을 대상으로 데이터세트의 기록화 연구가 수행된 바 있다. 각 논문에서는 행정정보시스템에 대한 특성을 분석하고 해당 업무시스템이 갖는 가치, 기록으로 보존할 대상 등에 초점을 맞추어 기록화 방안을 제시하고 있다. 국가적인 차원에서 데이터세트에 대한 기록관리 체계가 제시된 이후에도 개별 행정정보시스템에서 보존할 데이터세트 기록을 선별하고 획득하는 절차와 방법을 구체화하는 연구는 지속되어야 할 것이다. 행정정보시스템마다 역할이 다르고 데이터세트의 내용이 다르므로 선별 기준을 적용하는 기법이나 획득 방식의 특성이 다르기 때문이다. 다만 그간의 연구들은 데이터세트의 실질적인 이관 경험을 토대로 하고 있지 못하므로 이 논문에서 다루고자 하는 문제점의 인식에 도달하지 못하고 있다.

데이터세트 기록관리에 관한 연구의 두 번째 영역은 데이터세트 기록의 관리체계 전반에 관한 연구이다. 현문수[18]는 데이터세트에 대한 조직 및 기술에 대한 내용을 중심으로 제시하였고, 조은희 등[19]은 데이터세트를 기록으로 식별 및 선별하는 기준과 절차에 대한 연구를 수행한 바 있다. 또한, 국가기록원에서는 상기한 바와 같이 데이터세트 아카이빙을 위해 두 차례 연

으로」, 한국외국어대학교, 2007

14) 한철희, 「NEIS 교무업무시스템 데이터의 기록화 방안 연구(학교생활기록부를 중심으로)」, 명지대학교 기록과학대학원, 2007

15) 이은별, 「국가 재정정보의 기록학적 관리방안 : 디지털예산회계시스템을 중심으로」, 명지대학교 기록과학대학원, 2008

16) 이순한, 「조달업무의 설명책임성 확보를 위한 조달정보 기록관리 요건 연구」, 한국외국어대학교, 2008

17) 진채환, 「공공기관의 인사기록관리에 관한 연구」, 한국외국어대학교, 2007

18) 현문수, 「데이터세트 기록의 관리 방안」, 『한국기록학회지』, 2005

19) 조은희, 임진희, 「행정정보 데이터세트 기록의 선별 기준 및 절차 연구」, 『기록학연구』 제19호, 2009.

www.kci.go.kr

구 과제를 수행한 바 있다. 이들 연구는 데이터세트 아카이빙 체계를 갖추는 과정에서 필요한 개념, 프로세스, 시스템 등 기반 구축에 초점을 맞추고 있다. 데이터세트 기록의 이관에 관한 프로세스와 포맷 등에 관해서도 다루고 있으나 아직 본격적인 데이터세트의 이관을 실행한 경험이 없으므로 데이터의 값과 품질에서 발생할 수 있는 문제점과 이에 대한 대응 방법을 제시하지는 못하고 있다. 향후 아카이빙 프로세스 각 단계를 세분화하여 구체적인 주제 연구가 요구된다.

다음으로 전자기록 이관에 관한 선행연구를 살펴보고자 한다. 오삼균[20]은 사실상 국제 표준으로 인정받는 기준에 대한 분석과 사례연구를 토대로 이관 절차를 제안한 바 있다. 천권주[21]는 OAIS참조모형과 미항공우주국의 우주데이터시스템 위원회(CCSDS: Consiltative committee for Space Data System)에서 개발한 '생산자-아카이브 인터페이스 표준방법론'을 분석하고 해외의 이관 사례 고찰을 통해 이관절차모형을 제시하고 있다. 두 연구에서 제시하는 이관 모형은 데이터세트 아카이빙에도 적용될 수 있는 것으로, 전자기록을 이관하는 절차 중 '전송준비' 혹은 '이관준비' 단계에서 기록물의 오류를 검사하여 수정 보완하거나 보정하는 프로세스를 제시하고 있는 점은 이 논문의 주제와 연관하여 시사하는 바가 크다. 그러나, 선행연구는 기록물에 어떤 오류가 있을 수 있으며 수정 보완이나 보정을 어떻게 해야 하는지에 관한 구체적인 제시가 없다는 한계를 갖고 있다.

이상에서 살펴본 데이터세트 이관 시 품질에 관련한 연구를 종합하면, 전자기록의 이관 과정에서 품질 개선을 위해 기록의

20) 오삼균, 김희섭, 오상훈, 권도윤, 원선민,「전자기록물 이관 절차 개발에 관한 연구」, 한국문헌정보학회지, 2008.12

21) 천권주,「전자기록의 장기보존을 위한 이관절차모형에 관한 연구」, 기록학연구, 2007

www.kci.go.kr

수정 보완 절차가 필요하다는 것이 이관 모형에서 제시되어 있기는 하나 실제 데이터세트 이관 경험이 부재하므로 구체적인 문제 기반의 연구에는 이르지 못하고 있다. 따라서 이 논문에서 데이터세트 기록의 특성을 고려하여 이관 과정에서 발생할 수 있는 데이터 보정 및 품질 개선의 구체적인 사례를 다룸으로써 문제 기반의 구체적인 연구 성과가 될 수 있을 것이다.

이 논문에서는 데이터세트의 이관이 데이터웨어하우스의 ETT (Extraction, Transformation & Transportation, 추출, 변형 & 전송) 과정과 유사성을 가진다는 점에 착안하여 2장에서는 먼저 데이터웨어하우스 구축 과정에서 데이터를 보정하고 품질을 개선하는 절차와 방법에 대해 살펴보고자 한다. 또한, 데이터웨어하우스 구축과 아카이브 구축의 배경이 어떤 유사점과 차이점을 갖는지 살펴봄으로써 데이터세트 기록의 보정과 품질 개선의 특성을 이해하고자 한다. 3장에서는 행정정보 데이터세트에 예상되는 전형적인 데이터 보정 및 품질 개선 사례를 제시하고자 한다. 4장에서는 데이터 보정 및 품질 개선 내용이 행정정보 데이터세트의 생산시 품질요건으로 적용될 수 있음을 제안하고, 공공정보 활용을 위한 '민간 활용 지원센터' 설치에 따라 행정정보 데이터세트를 이관할 때도 유사한 데이터 보정 및 품질 개선 사례가 적용될 수 있음을 논의하고자 한다.

www.kci.go.kr

2. 데이터웨어하우스의 데이터 보정 방법

2.1 데이터웨어하우스 개요

매출이나 업무활동이 일정 규모 이상이 되는 기업이나 공공기관에서는 전사적 데이터웨어하우스(Enterprise Data Warehouse)[22]를 구축하는 경우가 많다. 데이터웨어하우스란, 업무 트랜잭션(transaction) 정보가 생성되는 여러 종류의 정보시스템으로부터 주기적으로 데이터를 추출하여 하나의 저장소(repository)에 누적하여 모은 후 목적에 맞춰 가공함으로써 의사결정에 필요한 중요한 정보를 도출하기 위해 구축하는 시스템이다. 예를 들어, 마케팅 영역에서는 데이터웨어하우스를 구축하여 중장기적인 고객의 소비 트렌드를 분석하거나, 고객군의 특성별 제품 선호도를 분석하거나, 계절적 수요와 지역적 수요를 분석함으로써 매출을 극대화하고 고객의 만족도를 높이는데 효과를 보고자 한다.

데이터웨어하우스 구축의 경험에서 발견한 중요한 공통적 사실이 있다면 업무를 수행하면서 정보시스템에 입력된 트랜잭션 정보가 기대보다 품질이 좋지 못하다는 것이다. 소위 데이터가 '깨끗하지 못한(dirty)' 상태이므로 이 트랜잭션 데이터를 있는 그대로 모아 분석을 하려고 하면 데이터 간에 상호 모순된 상태가 발생하거나 처리가 복잡하고 때론 불가능한 상태가 발생하더라

22) 80년대 중반 이후 사용자의 의사 결정에 도움을 주기 위해 IBM에서 처음으로 도입한 개념이다. 기간 시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환·관리하는 방식으로, 목적별 데이터를 비롯해 기업 활동 전반에 필요한 정보를 회사 전체 데이터베이스로 일원화한 방식으로 구성하여 데이터의 추출, 저장, 조회 등의 활동을 가능하게 한다.

www.kci.go.kr

는 것이다. 그래서 데이터웨어하우스 구축의 과정에는 데이터를 깨끗하게 만들어주는 '클렌징(cleansing)' 과정이 필수로 포함되어 있다. 즉, 업무트랜잭션 시스템에서 데이터를 추출하여 데이터웨어하우스 저장소에 데이터를 붓기 전에 품질을 보정하는 여러 조치를 취하는 것이다. 이러한 절차를 데이터웨어하우스 영역에서는 ETT 과정이라고 한다[23].

2.2 ETT 과정의 이해

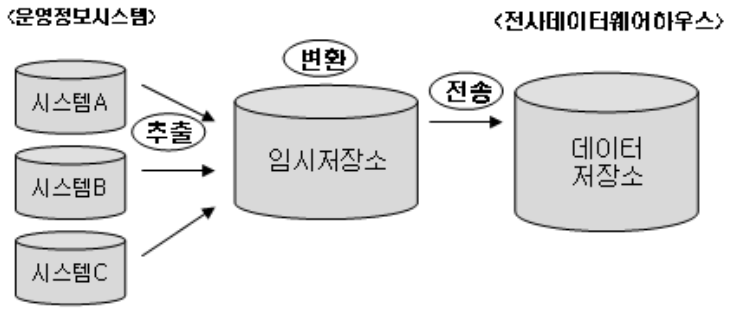

그림 1. 데이터웨어하우스의 ETT과정

전사적 데이터웨어하우스의 핵심 기술은 ETT를 효과적, 효율적으로 수행하는 것이다. 그림1에서 보는 것처럼 ETT는 조직에서 업무 수행에 사용하는 A, B, C 등 다수의 운영 정보시스템에 분산 저장되어 있는 각종 업무트랜잭션 처리 데이터를 목표 시스템인 데이터웨어하우스의 저장소로 취합하는 과정이다. ETT 과정을 통해 정확하고 명확한 자료를 추출함으로써 데이터웨어

23) ETT라는 용어 대신 ETL(Extract, Translation & Loading)이라고 부르는 등 유사한 용어들이 사용되고 있다.

www.kci.go.kr

하우스의 품질을 향상시킬 수 있다. ETT 방법은 데이터 출처인 운영정보시스템의 종류, 데이터의 추출주기, 데이터의 양, 로딩 속도, 출처 데이터의 질, 과거 데이터의 포맷, 사용자의 요구 조건, 출처 운영시스템의 역할 등에 따라서 달라진다. ETT의 각 단계에서 하는 작업 내용과 특징을 정리하면 다음과 같다.

첫째, ETT에서 추출(Extraction)은 운영정보시스템으로부터 데이터를 추출해 내는 과정이다. 운영정보시스템 별로 데이터를 추출하는 작업은 대부분 주기적으로 반복해서 이루어지기 때문에[24] 초기에 작업했던 방식을 잘 정리하고 원칙을 정해두면 이후의 추출 작업에 재활용할 수 있다. 이러한 점을 고려하여 추출 방법을 정책적으로 정하는 것이 장기적인 작업에 도움이 된다. 추출과정의 최종 작업은 데이터 품질(Data Quality)을 관리하기 위해 데이터 정제, 통합 규칙, 에러, 예외처리 규칙을 정의하는 것이다. 추출한 데이터를 임시 데이터베이스에 저장하여 상태를 확인하고 검토하면서 품질 관리를 위한 기준과 규칙 등을 정의한다. 둘째, ETT에서 변환(Transformation)은 추출한 데이터를 데이터웨어하우스에 축적하기 전에 규칙에 맞게 변환하는 것이다. 데이터 정제(refine)라고도 하며 데이터의 품질과 정확성을 향상시켜주는 단계이다. 출처 운영시스템에서 데이터를 추출하여 아무런 변환 없이 그대로 데이터웨어하우스에서 사용하는 경우는 거의 없다. 데이터베이스관리시스템(DBMS)의 종류가 다르기 때문에 포맷을 변경해야 하는 경우도 있고, 서로 다르게 구성되었지만 같은 의미를 가진 코드를 일치시키는 작업도 이루어지게 된다. 데이터웨어하우스로 여러 출처의 데이터가 합쳐지면

24) 조직이 의사결정을 하기 위한 정보를 제공해주는 목적으로 데이터웨어하우스를 구축하는 것이 일반적이다. 그러므로, 규칙적인 의사결정의 주기에 따라 데이터 추출이 반복적으로 이루어지게 된다. 예를 들어, 일별, 주별, 월별, 분기별, 반기별, 연도별 등.

www.kci.go.kr

서 조합해야 하는 과정인데, 합쳐진 데이터를 한꺼번에 조회, 연산할 수 있도록 정해놓은 규칙에 따라 변환하는 작업을 하게 된다. 셋째, ETT에서 전송(Transportation)은 변환된 데이터를 데이터웨어하우스에 전송하는 단계이다. 전송 방법으로는 온라인 방식과 오프라인 방식이 모두 가능하다.

ETT 과정에서 가장 중요한 이슈 중의 하나는 데이터의 정제 문제이다. 데이터 정제란 수집된 데이터를 분석에 활용하기 위해 일차적인 데이터의 결함을 제거하는 작업이다. 데이터웨어하우스는 다양한 시스템으로부터 많은 양의 데이터가 조합되는 곳이기 때문에 데이터의 오류나 이상이 발생할 가능성이 존재한다. 그래서 정제를 위한 도구를 사용하여 오류를 감지하고 교정하는 과정이 필요하다. 데이터의 정제가 필요한 전형적인 현상들은 다음과 같다[25].

- 동일한 내용의 필드 간에 길이가 서로 다른 경우
- 동일한 코드 값에 대해 설명 내용이 일치하지 않는 경우
- 동일한 필드에 서로 모순된 값이 들어있는 경우
- 값이 반드시 들어있어야 할 필드에 입력이 누락되어 있는 경우
- 무결성 제약조건에 위배되는 경우 : 예를 들어, 날짜 값이 유효 기간을 벗어나는 경우, 기본 키에 해당하는 필드의 값이 중복되는 경우 등

이러한 여러 변칙적인 상황(anomaly)을 해소하기 위해 조치들을 취하게 된다. 그림2는 이러한 목적으로 데이터베이스에서 추출한 데이터를 정제하는 과정을 제시하고 있다.

25) 이영재, 지능 의사결정지원시스템, 생능출판사, 2009

www.kci.go.kr

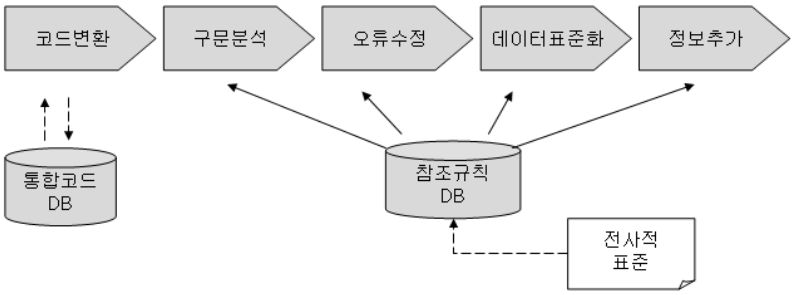

그림 2. ETT의 데이터 정제 과정

※ '이영재, 지능 의사결정지원시스템, 생능출판사, 2009'를 참조 변형하였음

첫 번째, 코드변환 작업은 출처 운영정보시스템에서 추출한 코드 값을 조직의 표준 코드에 맞춰서 변경해주는 조치이다. 예를 들어, 추출한 데이터에서는 남성을 '1', 여성을 '2'라는 코드로 가리키고 있고 조직의 표준 코드는 '남', '여'라고 한다면 표준 코드에 맞춰 값을 변경해준다. 이 때, 일반적으로 조직단위의 '통합코드DB'에 공통적으로 사용되는 코드값을 모아두고 이 값을 참조하여 사용하도록 한다. 만약 새로운 코드값이 처음으로 추출된 경우에는 '통합코드DB'에 해당 코드 값과 변환 값을 추가 등록하여 이후의 추출과정에서 참조하도록 한다.

두 번째, 구문분석 작업은 추출한 데이터가 여러 의미정보의 복합체로 구성된 경우 이를 각 의미단위로 분해하는 조치이다. 예를 들어, 주소값의 경우 시도 정보, 동 정보, 번지 정보, 우편번호 정보 등으로 구성되어 있는 하나의 데이터로 추출될 수 있다. 이 때, 지역별 고객들의 소비패턴이나 수입의 수준을 비교분석하는 마케팅 데이터웨어하우스를 구축하고자 한다면 주소값을 하나의 칼럼으로 저장하기 보다는 하위의 의미정보 단위

www.kci.go.kr

인 시도 정보, 혹은 동 정보를 추출해 내어 별도의 칼럼으로 저장하는 것이 유리하다. 이렇게 하위 의미정보를 추출하기 위해서는 데이터의 구문분석 작업이 이루어져야 한다.

세 번째, 오류수정 작업은 추출한 데이터의 값이 잘못된 값일 경우 이를 수정하여 바로잡는 조치이다. 예를 들어, 우편번호는 주소에 따라 정해지는 값이라는 것을 알고 있다면, 추출된 데이터에 주소와 우편번호가 서로 일치하는지 검증할 필요가 있고, 만약 서로 일치하지 않는 값이 발견된다면 이를 수정해주어야 한다는 것이다. 이를 위해서는 데이터 값을 어떤 규칙에 의해 검증할 것인지, 또한 어떤 규칙에 따라 값을 수정할 것인지를 정해야 한다. 예를 들어, 주소와 우편번호가 불일치할 때는 주소에 맞춰 우편번호를 수정한다, 는 식의 규칙을 정하는 것이다.

네 번째, 데이터 표준화 작업은 데이터 값을 표현하는 방식에 규칙을 두어 일관되게 하는 조치이다. 예를 들어, 날짜 값은 모두 'YYYYMMDD'[26]로 표시하고, 영문 주소명에서 '연희동'은 'Yeonhui-dong'으로 통일하여 표기한다고 규칙을 정하는 것이다.

다섯 번째, 정보추가 작업은 데이터의 의미를 해석할 수 있도록 필드를 추가해 주는 조치이다. 예를 들어, 출처 운영정보 시스템에서는 어느 부서인지 쉽게 알 수 있는 부서 코드값이 데이터웨어하우스로 옮겨졌을 때는 부서명을 알기 어려울 수 있다. 이러한 문제를 방지하기 위해 부서코드 값에 따른 부서명 필드를 추가할 수 있다. 이상에서 살펴본 구문분석, 오류수정, 데이터 표준화, 정보추가 등의 작업을 할 때는 참조규칙DB에 저장된 규칙에 따르도록 한다. 참조규칙DB는 조직 전체의 표준

26) 'YYYYMMDD'에서 YYYY는 연도 4자리, MM은 월 2자리, DD는 일 2자리를 의미한다. 즉, 2010년 4월 1일을 '20100401'로 표현하는 것이다.

www.kci.go.kr

을 준수하면서 다양한 사례에 맞춰 정제 규칙을 축적하여 만들어 가는 규칙 데이터베이스이다.

2.3 데이터웨어하우스와 데이터세트 아카이브 구축의 비교

데이터세트를 아카이브로 이관하는 것은 데이터웨어하우스를 구축하는 것과는 목적이나 배경, 작업환경 측면에서 이질적인 작업이다. 하지만, 여러 개의 출처 정보시스템에 있는 데이터베이스에서 필요한 데이터를 추출하여 하나의 통합된 데이터베이스 시스템으로 집적하여 관리한다는 측면에서 구조적, 기술적 유사성을 갖는다. 이러한 유사성에 근거하여 데이터웨어하우스의 ETT과정을 기반으로 아카이브로 입수되는 데이터세트의 보정 및 품질 개선에 대한 시사점을 도출할 수 있다. 정확한 시사점을 얻기 위해 먼저 데이터웨어하우스와 데이터세트 아카이브 구축과정의 특징을 비교해보고자 한다.

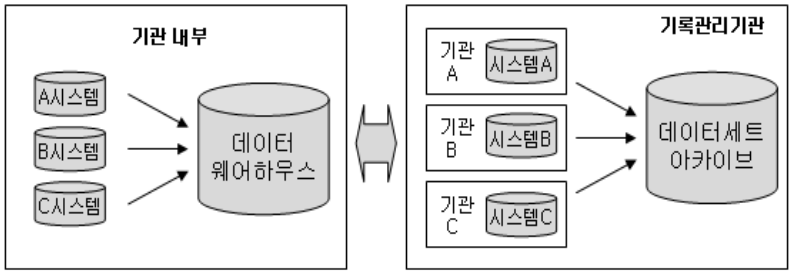

그림 3. 데이터웨어하우스와 데이터세트 아카이브의 비교

데이터웨어하우스와 데이터세트 아카이브는 구축의 기본 목적과 배경이 다르므로 다음과 같은 차이점을 갖는다. 첫째, 그

www.kci.go.kr

림3에서 보는 것처럼 데이터웨어하우스의 구축은 대부분 기관 내의 데이터 이동을 전제로 하지만 데이터세트 아카이브는 기관 간의 데이터 이동을 전제로 한다. 데이터웨어하우스의 저장소가 분산환경의 네트워크 상에 존재하더라도 동일한 조직의 관할권 내에 존재하게 된다. 이처럼 저장소가 조직 내부의 시스템인 경우 데이터베이스 트리거를 작성하여 ETT 과정을 처리하는 것이 가능해진다. 반면 공공 행정정보 데이터세트 아카이브의 경우 데이터가 정부기관에서 국가기록원과 같은 기록관리기관으로 이동하게 되므로 출처 데이터시스템과 아카이브시스템의 관할권이 달라진다. 공식적인 데이터세트 이관 후 출처 데이터시스템의 상황 정보는 차단될 가능성이 많으므로 이관되는 데이터세트에 대한 결함이나 애매모호한 부분이 존재하지 않도록 처리하는 것이 필요하다.

둘째, 데이터웨어하우스와 데이터세트 아카이브는 사용자 집단과 기본적인 요구사항이 다르다. 데이터웨어하우스는 기관 내부 사용자를 대상으로 하지만 데이터세트는 기관 외부 사용자를 대상으로 한다. 따라서 데이터세트의 경우 업무맥락에 익숙하지 않은 사용자들도 내용을 이해할 수 있도록 데이터항목에 관해 충분한 설명을 부과하는 것이 필요하다. 또한 기관 외부로 공개 시 데이터세트에 관한 저작권이나 개인정보 보호 문제가 발생하지 않도록 유의해야 한다.

셋째, 앞에서 살펴본 차이점으로 인해 데이터웨어하우스와 데이터세트 아카이브는 데이터 집적 과정을 시스템적으로 자동화하는 방식과 범위에 차이가 있다. 데이터웨어하우스는 대부분 시스템 자동화가 가능하고 이를 근본적으로 지향하지만 데이터세트 아카이브는 데이터세트를 제공하는 기관의 상황에 따라 편차가 클 수 있다.

www.kci.go.kr

이러한 차이점에도 불구하고 데이터웨어하우스와 데이터세트 아카이브 구축과정은 다음과 같은 유사성을 갖는다. 먼저, 둘 다 데이터 정제 과정이 필요하다는 점이다. 출처 정보시스템의 데이터와 목표 정보시스템은 운영 목표가 다르므로 데이터의 품질에 대한 관심과 기대수준이 다르다. 예를 들어, 운영 상에서는 일부 모순된 데이터가 업무에 큰 지장을 초래하지 않을 수 있지만 데이터웨어하우스나 아카이브로 옮겨왔을 때는 데이터 간의 상호 정합성과 무결성이 중요하므로 문제가 될 수 있는 것이다. 따라서, 목표시스템의 목적에 맞춰 일정 수준으로 데이터를 정제할 필요가 있다. 둘째, ETT에 필요한 기술적 기반이 유사하다는 점이다. 출처 정보시스템에서 데이터를 추출하여, 변형하고, 목표 정보시스템에 적재하는 데는 ETT 작업을 위한 별도의 데이터 저장 공간과 도구가 필요하다. 또한, 다중의 출처로부터 데이터를 통합하여 저장하기 위해서는 분산데이터베이스 기술을 적용하는 것이 필요하다. ETT 작업이 주기적, 반복적으로 실행되어야 할 때는 추출 및 변형 로직을 만들어 자동화시킬 필요가 있다. 셋째, 기관의 표준화된 규칙이 정해져 있어야 한다. 데이터에 대한 정제가 필요할 때 어떤 규칙을 기반으로 작업할 것인지를 정하는 것이 필요하다. 코드값을 하나로 일치시킨다거나 변환의 방식을 정의하는 등 최소한 기관 단위에서 혹은 정부 차원에서 표준적 기준을 만들어 준수해야 한다.

데이터세트 기록을 아카이브로 입수하는 과정은 그림4와 같이 데이터세트 생산기관과 아카이브 기관 간에 여러 단계에 걸친 협상과 조치로 이루어진다.

www.kci.go.kr

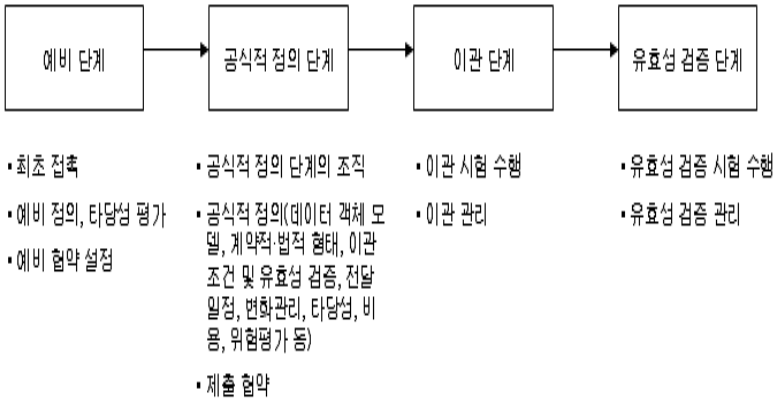

그림 4. 생산자와 아카이브간의 정보를 전송하는 프로세스

※ 인용 : Producer-Archive Interface Methodology Abstract Standard, Recommendation for Space Data System Standards, CCSDS 650.0-B-1. Blue Book. Issue 1, 2004

각 단계의 데이터 보정과 품질 개선 과제를 정의해보면 다음과 같다. 예비 단계에서는 생산기관과 아카이브가 데이터세트의 보정 및 품질 개선에 관한 일반적인 원칙에 합의하고, 공식적 정의 단계에서 세부적인 데이터세트 보정 및 품질 개선 요건을 도출하고 합의하는 것이 필요하다. 이관 단계에서는 이관대상 데이터세트 기록을 실제로 추출하여 보정 및 품질 개선을 실행하며 그 결과 데이터세트를 아카이브로 저장한다. 유효성 검증 단계에서는 데이터세트 기록의 진본성이 유지된 상태로 데이터 보정 및 품질 개선이 효과적으로 완료되었는지 확인한다.

www.kci.go.kr

단계 / 태스크	예비	공식	이관	유효성 검 증
수행 태스크	최초접촉 예비정의, 타당성 평가 예비 협약 설정	단계의 조직 여러가지에 대한 공식적 정의 제출 협약	이관 시험 수행 이관관리	유효성 검증 시험 수행 유효성 검증 관리
데이터 보정 및 품질 개선 태스크	보정과 개선의 원칙 합의	보정과 개선의 요건 도출 과 합의	보정 및 품질 개선 실행	보정과 개선 효과 확인

표 1. 데이터세트 이관 시 단계별 태스크

3. 데이터세트 기록 이관 시 데이터 보정 및 품질 개선 사례

이 장에서는 그림2에 제시한 데이터웨어하우스의 데이터 정제 과정 별로 행정정보 데이터세트 이관 시 데이터 보정 및 품질 개선의 사례를 살펴보고자 한다.

3.1 추출(extraction) - 데이터세트의 수량과 유효값 확인

이관을 위해 추출할 데이터 집합에 대해 수량을 확인하는 것은 기본적인 품질확인 절차이다. 이는 실재하는 데이터의 집합과 응용프로그램을 통해서 확인한 데이터의 집합에 차이가 있을 수 있기 때문이다. 다음으로는 추출 대상 데이터 집합의 각 필드들이 유효한 값인지를 검토해야 한다.

예를 들어 이런 사례가 있을 수 있다. 데이터베이스 상의 테

www.kci.go.kr

이블에서 직접 데이터세트를 질의(query)하여 추출해 보면 100건의 데이터 건이 선별되는데, 해당 데이터세트를 처리하는데 사용하는 주요 응용프로그램을 통해 데이터를 조회하면 98건만 선별되는 경우이다. 이는 테이블에는 실제 100건의 데이터가 입력되어 있으나 상태(status)가 '무효(invalid)'한 데이터가 2건이 있었고 그림5와 같은 구조에서 응용프로그램의 로직을 통해이를 걸러 주기 때문이다[27]. 즉, 행정정보시스템의 사용자가 화면상에서 질의하거나 보고서를 출력했을 때 확인할 수 있는 데이터의 수량과 직접 데이터베이스에서 질의한 데이터가 같지 않을 수 있다는 것이다.

그림 5. 응용프로그램을 통해 데이터를 확인하는 사용자

이처럼 특히 응용프로그램을 통해 생성되고 축적된 데이터세트의 경우는 직접 데이터베이스를 질의하여 추출한 데이터세트가 응용프로그램을 통해 접근할 수 있는 데이터세트와 다를 수 있다는 점에 유의해야 한다. 이는 수량의 측면뿐만 아니라 필드가 유효한 값을 갖고 있는지를 확인할 때도 적용된다.

응용프로그램은 데이터베이스의 값을 읽어서 화면에 디스플레이하거나 보고서로 출력해줄 때 필드 값을 해석하거나 가공

27) 질의문인 select 문장에 where절에서 상태값 등을 확인하여 로우(row)를 걸러주는 경우가 많다.

www.kci.go.kr

하는 로직을 포함하기도 한다[28]. 예를 들어, 성별을 의미하는 필드의 값이 1 이면 '남성', 2 면 '여성'으로 해석하여 화면과 보고서에는 1 혹은 2라는 필드 값 대신에 '남성' 혹은 '여성'이라는 값을 보여주는 식이다. 그런데, 만약 실수로 필드 값이 1이나 2가 아닌 다른 값, 즉 0과 같은 입력되어 있는 경우에는 응용프로그램의 로직에서 예외처리를 하여 오류표시를 하는 등 별도의 처리를 하여 보여주게 될 것이다. 이 경우 데이터베이스의 성별 필드 값을 직접 추출하게 된다면 1이나 2 이외의 값이 그대로 포함되므로 응용프로그램을 통해 확인하는 것과는 다른 상태의 데이터세트를 얻게되는 것이다.

위와 같은 점들을 고려할 때, 데이터세트를 추출하는 루틴을 작성할 때 데이터베이스에 직접 질의하는 방식보다는 행정정보시스템의 주요 업무 메뉴에서 업무담당자가 화면과 보고서로 유효한 집합의 값을 확인한 후 바로 익스포트(export)[29]하는 방식이 적합하다고 볼 수 있다. 물론, 이 때 화면에는 나타나지 않지만 데이터베이스에 물리적으로 존재하는 여러 칼럼들 중 어디까지를 함께 익스포트할 것인지를 선택할 수 있는 화면이나 도구를 별도로 제공하면 편리할 것이다.

28) 오라클 데이터베이스의 경우 decode 함수를 사용하여 필드 값을 해석하여 처리하는 경우가 많다.

29) 예를 들어, 조회한 데이터세트를 엑셀과 같은 스프레드시트 파일로 만들어주는 기능이다.

www.kci.go.kr

3.2 코드 변환 - 일관된 코드값 부여

Table 1	Table 2	Table 3
gender1	gender2	gender3
남	M	01
여	F	02
여	F	02
남	M	01

데이터의 저장 상태

Table 1	Table 2	Table 3
gender1	gender2	gender3
M	M	M
F	F	F
F	F	F
M	M	M

코드변환 후 데이터 상태

그림 6. 성별 코드 값을 일치시킨 사례

데이터세트에 포함된 특정 정보가 코드화되어 관리되는 경우가 많다. 예를 들어, 성별 정보의 경우, '남자', '여자' 와 같은 값을 갖게 되는데 데이터베이스에 정보를 저장할 때는 그림6과 같이 다양한 형태의 코드값을 부여하여 사용하는 경우가 많다. 그림6의 왼쪽 테이블을 보면, Table1의 경우 '남', '여'라는 값으로, Table2의 경우 'M', 'F'라는 값으로, Table3의 경우 '01','02'라는 값으로 각각 '남자'와 '여자'라는 값을 대표한다. 이러한 코드값을 포함하는 데이터세트를 추출할 때는 다른 데이터세트와의 일관성을 고려할 필요가 있다. 실제로는 같은 의미를 가진 정보가 서로 다른 코드 값으로 전송되었을 때 이후 과정에서 의미 해석에 혼란이 있을 수 있기 때문이다. 그림6의 예시에서 각 테이블이 서로 다른 데이터세트에 포함되어 추출되는 경우라면, 기관 레벨에서 혹은 아카이브 레벨에서 코드 값을 통일하도록 협의가 이루어져야 하며 하나의 대표 코드 값을 선정하여 나머지 데이터세트에서는 코드 값을 변경하여 일치시켜주도록 한

www.kci.go.kr

다. 만약 성별 코드를 'M','F'로 통일시키기로 한다면 그림6의 오른쪽과 같이 추출한 성별 코드 값을 변환시켜줄 수 있다. 그림2에서 제시한 바와 같이 기관에서는 조직 차원에서 혹은 정부기관 전체의 차원에서 통합코드DB를 만들어 운영하는 것이 효과적이다.

3.3 구문분석 - 복합정보의 컴포넌트화

Table 1
ref_code
1234567-AA00120-000001
1234567-AA00120-000002
1234567-AA00120-000003
1234567-AA00120-000004

Table 1
ref_code
처리과코드+ 단위업무코드+ 일련번호
처리과코드+ 단위업무코드+ 일련번호
처리과코드+ 단위업무코드+ 일련번호
처리과코드+ 단위업무코드+ 일련번호

그림 7. 복합정보로 이루어진 코드 값 사례

그림7에서 Table1의 Ref_code 칼럼 값은 처리과코드와 단위업무코드값을 조합하고 여기에 일련번호를 붙인 복합정보이다. 데이터베이스의 테이블에는 데이터 건들을 서로 구별하기 위해 기본키를 정의하여 사용하는 것이 일반적이다. 이 때, 식별자 역할을 하는 값으로 각종 코드값을 정의하게 되는데, 여러 개의 코드를 조합하여 키를 생성하는 경우 각각을 칼럼으로 설계할 수도 있고, 그림7과 같이 하나의 복합 키 칼럼으로 설계할 수도 있다. 만약 이 데이터를 아카이브로 이관하여 서비스한다면 사용자들이 Ref_code 값을 이해하기는 어려울 것이다. 따라서, 다음과 같은 점들을 검토하여 추출 방식을 결정해야 할 것이다.

www.kci.go.kr

먼저, Ref_code 값이 응용프로그램에서 어떻게 사용되고 있는지를 살펴보아야 한다. 데이터베이스 상에는 칼럼이 하나이지만 사용자 뷰에 디스플레이할 때는 3개의 하위 의미정보별로 나누어 보여주거나 의미정보별로 각각 다른 테이블들과 조인하기 위해 사용되고 있을 수 있다. 예를 들어, 처리과코드만 추출하여 조직정보와 조인하고, 단위업무코드만 추출하여 업무분류정보와 조인하는 방식이다. 만약 사용자 뷰에서는 하나의 전체 덩어리로 쓰이기 보다는 의미있는 조각으로 나누어 쓰는 경우도 많다면 해당 데이터를 추출하여 아카이브에 입수할 때 하위 의미정보별로 칼럼을 분해하는 것에 대해 결정해야 한다. 또한, 3.2에서 살펴보았듯이 코드값에 해당하는 설명정보도 함께 포함할 것인지 여부까지 검토해야 한다.

Table 1
ref_code
처리과코드+ 단위업무코드+ 일련번호
처리과코드+ 단위업무코드+ 일련번호
처리과코드+ 단위업무코드+ 일련번호

대안1. 복합정보 그대로

Table 1		
ref_code1	ref_code2	ref_code3
처리과코드	단위업무코드	일련번호
처리과코드	단위업무코드	일련번호
처리과코드	단위업무코드	일련번호

대안2. 하위 의미정보로 분해

Table 1				
ref_code1	ref_code1_desc	ref_code2	ref_code2_desc	ref_code3
처리과코드	처리과코드명	단위업무코드	단위업무명	일련번호
처리과코드	처리과코드명	단위업무코드	단위업무명	일련번호
처리과코드	처리과코드명	단위업무코드	단위업무명	일련번호

대안3. 의미정보로 분해하고 설명정보 추가

그림 8. Ref__code 값을 추출할 때의 가능한 대안

www.kci.go.kr

그림8은 Ref_code 값을 추출할 때의 가능한 대안들을 보여준다. 대안1과 같이 복합키 값을 그대로 전송해 가는 경우, 대안2.와 같이 3개의 하위 의미정보로 분해하여 전송해 가는 경우, 대안3.과 같이 3개 하위 의미정보로 분해하고 여기에 각각의 설명정보까지 붙여 전송해 가는 경우가 있을 수 있다. 이 때, 그림2에서 제시한 참조정보DB에 복합정보를 처리하는 규칙을 저장해 두고 이를 참조하여 일관되게 처리하는 것이 효과적이다.

3.4 오류수정 - 날짜 데이터의 정밀도 결정

Table 1
create_date
2009010000000
20071215103131
20050430000000
20050430103131

응용프로그램의 화면
2009년 1월 3일
2007년 12월 15일
2005년 4월 30일
2005년 4월 30일

그림 9. 저장된 데이터와 사용되는 데이터의 정밀도가 다른 경우

데이터세트는 이를 생성하고 검색하는데 필요한 응용프로그램과 한 쌍으로 존재하는 경우가 많다. 그림5와 같이 사용자와 데이터베이스 사이에서 응용프로그램은 데이터를 생성, 조작, 삭제해주며 사용자 맞춤 형태로 데이터를 디스플레이 해준다. 사용자 뷰에서 보는 데이터의 형태와 실제 데이터베이스에 존재하는 데이터 형태에는 일정한 격차가 존재하며 데이터를 조작하는 전 과정에서 데이터의 변형이 지속적으로 이루어지고 있는 것이다. 따라서 데이터세트를 추출할 때 데이터를 어떻게

www.kci.go.kr

변형하여 가져올 것인지를 정의하는 일이 중요하다. 그 예로서 날짜 데이터의 변형 사례를 살펴보고자 한다.

업무 트랜잭션을 처리하는 정보시스템에서는 하나의 트랜잭션 당 여러 개의 날짜 데이터가 발생하는 것이 일반적이다. 민원처리시스템의 경우, 민원인이 민원을 제기한 날짜, 민원상담실에서 확인하여 접수한 날짜, 예정 처리 날짜, 민원 처리 결과를 올린 날짜, 이의를 제기한 날짜, 그에 대한 답변을 올린 날짜 등 민원처리 건의 상태가 변화될 때마다 날짜 값을 따로 저장해 둔다. 이러한 날짜 값이 포함된 데이터세트를 추출하고자 할 때는 다음과 같은 점에 유의해야 한다.

데이터베이스에 날짜 타입의 칼럼이 정의되어 사용된다면 오라클 데이터베이스의 경우 세기구분, 연도, 월, 일, 시, 분, 초의 값이 7바이트 길이의 필드 안에 독자적(proprietary) 포맷으로 인코딩되어 들어간다. 오라클 데이터베이스에서 날짜값을 읽어 화면에 디스플레이하는 응용프로그램은 필요에 따라 연도와 월, 일을 추출하거나 혹은 시, 분, 초를 추출하게 된다. 그림9는 Table1의 create_date 칼럼에 연,월,일,시,분,초 정보가 모두 들어가 있는 경우와 시,분,초 정보는 0으로 초기화된 경우가 섞여있는 저장된 경우를 보여주고 있다. 응용프로그램은 해당 칼럼을 읽어서 화면에 표시해 줄 때는 '2009년 1월 3일', '2007년 12월 15일' 등과 같이 형식을 갖추어 디스플레이 해주고 있다. 이 업무에서의 전제는 연도, 월, 일 까지만 유의미한 정보였던 것이다. 이 경우, create_date 칼럼의 값이 시, 분, 초의 정보를 포함하느냐에 관계없이 응용프로그램이 처리한 결과는 일관될 것이다. 그러나, 이런 상태의 create_date 칼럼 값을 데이터세트로 이관하게 되는 경우에는 혼란이 예상된다. 아카이브에서 서비스할 때 날짜 값의 어디까지가 유의미한 값인지를 알지 못한 채 전체 날짜

www.kci.go.kr

값을 서비스한다면 사용자들은 통제되지 않은 시, 분, 초 정보에 접근하게 될 것이기 때문이다. 이러한 오류를 방지하기 위해서는 유의미한 날짜 정보를 제외한 부분의 값은 무시할 수 있어야 할 것이다.

위와 같은 점들을 고려할 때, 날짜 값 데이터를 추출할 때는 다음의 사항을 검토하여 변형을 결정해야 한다. 먼저, 날짜 값 칼럼에서 관리되는 정보의 종류를 확인한다. 예를 들어, 세기, 연, 월, 일, 시, 분, 초, 혹은 더 세분화된 수준의 타임스탬프 중 어디까지 저장되는 지를 살펴본다. 둘째, 유의미하게 관리하는 날짜 값 정보의 종류 및 수준을 확인한다. 응용프로그램에서 입력을 받을 때와 출력해줄 때의 포맷을 확인하면 알 수 있다. 셋째, 실제 데이터베이스에 저장된 날짜 값의 패턴을 검토한다. 만약 유미하게 관리하는 정보 이외의 더미 정보가 발견된다면 이를 어떻게 변형 처리할 것인지 결정한다. 넷째, 데이터 추출 루틴에 변형 로직을 적용하여[30] 무시할 정보 값들 대신에 정해진 디폴트 값이나 0을 추출할 수 있도록 조치한다.

3.5 데이터 표준화 - 시점 정보의 표준시간대 적용

2010년 4월 1일 오전 10시 정각이라는 시각은 한국에서와 미국에서가 전혀 다른 시점이 된다. 시각 정보는 시간대(time zone) 정보가 합쳐져야 유일한 시점을 확인할 수 있다. 글로벌 환경에서 진행되는 트랜잭션 데이터세트에는 여러 나라, 서로 다른 시간대에서 취해진 업무활동 내역이 혼재되어 있을 수 있다. 이 경우, 날짜 값이나 타임스탬프 값을 입력한 지역의 시간대를 감

30) 오라클의 경우 round, trunk 등의 함수를 활용하여 매년 1월 1일 날짜 값이나, 혹은 매월 1일 날짜 값, 혹은 매일의 0시 날짜 값 등을 계산해 낼 수 있다.

www.kci.go.kr

안하여 해석하지 않으면 정확한 시점을 파악할 수 없게 된다.

이러한 점을 고려할 때, 날짜 값이나 타임스탬프 값을 추출할 때는 다음의 사항을 검토하여 변형을 결정해야 한다. 먼저, 날짜 값이나 타임스탬프 등의 시각 정보가 시간대 정보를 함께 포함하고 있는지 확인한다. 혹은 표준화된 시점정보를 사용하고 있는지를 확인한다. 둘째, 시간대 정보를 포함하지 않은 시각정보인 경우, 시간대를 특정한다. 셋째, 데이터를 표준화된 시점정보로 추출할 것인지, 특정 시간대를 포함한 시각정보로 추출할 것인지 기준을 정한다. 넷째, 데이터 추출 루틴에 원하는 시점정보 작성을 위한 변형 로직을 적용한다.

3.6 정보추가 - 코드값의 설명정보

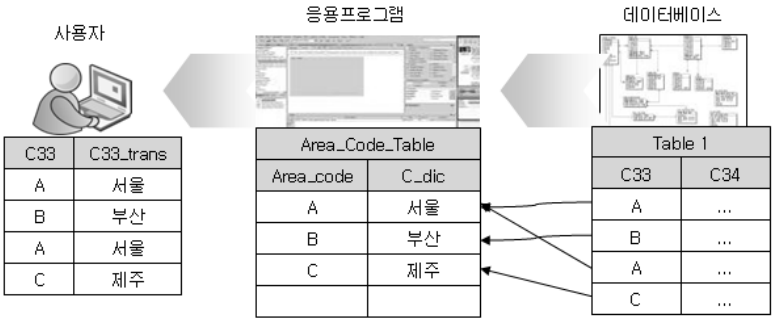

그림 10. 응용프로그램에서 코드 설명정보를 가진 경우

행정정보시스템의 데이터베이스에는 다양한 코드값이 사용되고 있을 것이다. 지역코드, 성별코드, 직업군코드, 산업군코드, 업종코드, 기관코드 등 수많은 코드값들은 숫자나 문자의 조합으로 이루어지는 경우가 많으며 그 자체로는 의미를 포함하지

www.kci.go.kr

않는다. 따라서 코드값에는 이를 설명해주는 정보가 필요하다. 일반적으로 코드값의 설명정보를 제공하는 방식은 다음 세 가지 방식 중의 하나이다.

첫째, 데이터베이스에는 코드값만 저장하고 있으면서 응용프로그램에서 해당 코드값을 화면이나 보고서에 보여줄 때는 설명정보를 별도로 제공하는 방식이다. 그림10에서 보는 Table1처럼 C33 칼럼은 코드값만 저장하고 있고, 이 값을 읽어들인 응용프로그램의 로직에서 C33의 값이 'A'이면 '서울', 'B'이면 '부산', 'C'이면 '제주'와 같이 데이터베이스의 값에 따라 설명정보를 붙여주는 방식이다.

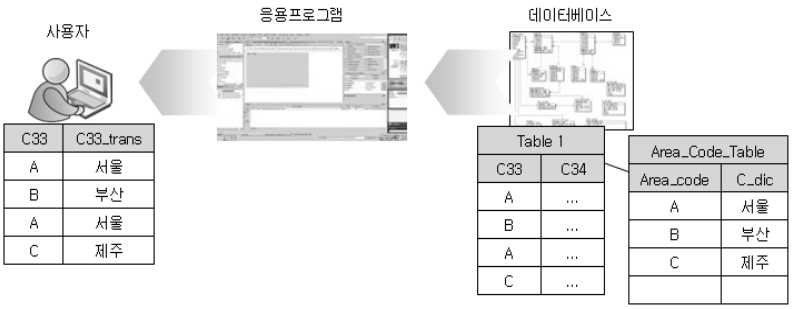

그림 11. 데이터베이스에 코드테이블을 가진 경우

둘째, 데이터베이스에 따로 코드테이블을 두고 응용프로그램에서 코드값을 읽을 때는 항상 코드테이블과 조인하여 설명정보를 함께 읽어와 제공하는 방식이다. 그림11에서 보는 Table1의 C33 칼럼에는 코드값이 저장되어 있고, 별도의 Area_code_table에 코드값의 설명값인 area_name 칼럼이 존재하는 것을 전제로 한다. 이 경우 응용프로그램에서 일관되게 Table1의 C33과 Area_code_table의 area_code 칼럼을 조인하여 area_name 값을 참조하

www.kci.go.kr

는 방식이다.

셋째, 외부 데이터베이스에 따로 코드테이블을 두고 응용프로그램에서 코드값을 읽을 때 항상 외부 코드테이블을 참조하여 설명정보를 제공하는 방식이다. 정부기관의 경우 정부기관 코드나 BRM 업무기능코드 등 별도의 정보시스템에서 코드값을 제공하는 사례가 적지 않다.[31] 그림12에서 보듯이 응용프로그램에서 Table1의 C33과 디렉토리서버나 공유서버의 Area_code_table의 area_code 칼럼을 조인하여 area_name 값을 참조하는 방식이다.

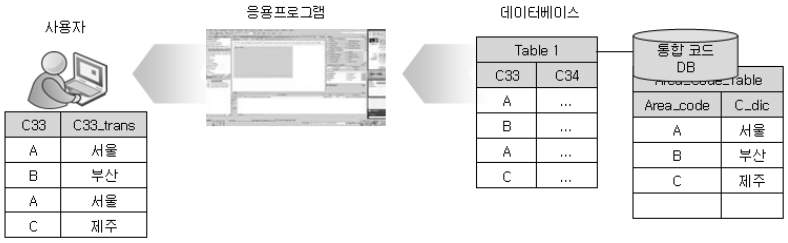

그림 12. 통합코드 DB에 코드테이블을 가진 경우

위와 같은 점들을 고려할 때, 추출하는 데이터세트에 코드값이 존재하는 경우 코드값의 의미를 설명해주는 정보가 포함되었는지를 확인하는 것이 필요하며, 데이터세트를 추출하는 루틴을 작성할 때 코드테이블을 존재하는 경우 조인하여 설명정보를 함께 가져오고, 코드테이블이 존재하지 않는 경우 응용프로그램의 로직을 조사하여 설명정보를 파악하여 함께 가져오도록 할 필요가 있다. 향후 데이터세트를 활용 및 서비스할 때 사

31) 전자정부사업 추진시 각급기관의 정보자원의 원활한 공유 및 공통서비스를 지원하기 위해 행정표준코드관리시스템(https://code.gcc.go.kr/), 정부디렉토리시스템(https://www.dir.go.kr/) 등을 운영하고 있다.

www.kci.go.kr

용자가 코드값을 해석할 수 있는 정보가 필요하기 때문이다.

3.7 정보추가 - 메타데이터 확보

기록으로서의 데이터세트는 아카이브에 입수되어 관리, 제공되는 전 과정에서 해당 데이터세트를 계층적으로 설명해주는 메타데이터가 필요하다[32]. 예를 들어, 아카이브가 행정정보시스템 - 데이터세트 시리즈 - 데이터세트 아이템 - 테이블 - 칼럼의 계층구조를 가지며, 다음과 같은 필수 메타데이터를 관리한다고 가정해보자.

- 행정정보시스템 레벨 : 해당 시스템명, 시스템 개통일자, 시스템 소유기관코드/명, 사용위치 혹은 주소, 소프트웨어와 하드웨어 명세 등과 같은 기술(description) 메타데이터
- 데이터세트 시리즈 레벨 : 해당 시리즈명, 내용 설명, 데이터베이스명, 데이터베이스관리시스템명, 추출 주기/추출방법, 데이터 언어, 아이템 시작일과 마지막일 등
- 데이터세트 아이템 레벨 : 내용 설명, 추출일자, 추출작업자, 데이터 바이트 수 등
- 테이블 레벨 : 테이블명, 내용 설명, 칼럼 수, 생성일자, 데이터 건수, 테이블 소유자 등
- 칼럼 레벨 : 칼럼명, 내용 설명, 데이터 유형, 유효한 값의 범위 등

32) 데이터세트의 관리 계층을 설정하는 이유는 업무적 연관성이나 생애주기에 따라 통제가 효율적으로 가능하고, 이용자가 기록을 이해하는데 도움을 주기 위해서이다. 조은희, 임진희, 「행정정보 데이터세트 기록의 선별 기준 및 절차 연구」, 『기록학연구』 제19호, 2009.

www.kci.go.kr

그림13에서 보는 바와 같이 메타데이터 값들은 데이터세트를 추출하는 과정에서 최대한 자동으로 추출하는 것이 효과적, 효율적이다. 특히, 행정정보시스템에서 보유한 정보거나 데이터베이스에서 보유한 정보의 경우 필수적으로 확보하여 함께 추출하는 것이 필요하다.

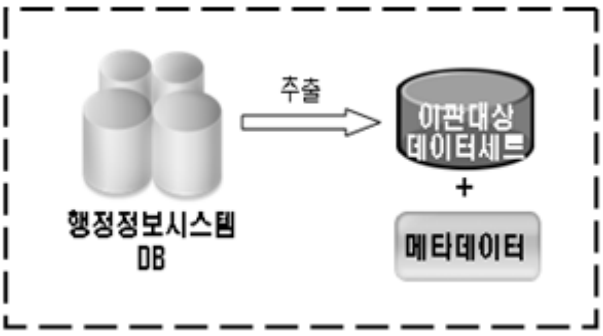

그림 13. 데이터세트 이관 시 메타데이터의 확보

예를 들어, 칼럼명, 칼럼의 데이터 유형, 테이블명, 테이블 칼럼 수와 생성일자, 데이터 건수, 소유자 등의 정보는 데이터베이스의 딕셔너리[33]에서 추출할 수 있다. 데이터세트의 추출일자, 추출작업자, 데이터 바이트 수, 데이터세트 시리즈의 데이터베이스명, 데이터베이스관리시스템명 등은 추출작업을 수행한 행정정보시스템에서 함께 추출하는 것이 필요하다. 데이터세트 이관작업자나 정보화담당자, 나아가 아카이브의 아키비스트가 직접 기술하는 노력과 부담을 경감시켜줄 수 있기 때문이다.

33) 오라클 데이터베이스의 경우 데이터 딕셔너리(dictionary)를 생성하여 관리하는데 딕셔너리는 데이터베이스에 생성된 객체들에 관한 메타데이터 집합이다.

www.kci.go.kr

4. 맺음말

데이터세트의 이관 시 적용하는 데이터 보정 및 품질 개선 사항 중 일부는 데이터세트를 생산하는 행정정보시스템에 대한 기본 데이터 품질요건으로 적용할 수 있다. 데이터세트를 주기적으로 이관해야 하는 행정정보시스템의 경우 데이터의 표준화와 품질 개선을 앞서 적용함으로써 이관 시 반복되는 보정과 개선 작업을 줄일 수 있다. 3장에서 다룬 데이터 보정 및 품질 개선 사례를 행정정보시스템의 데이터세트 품질요건으로 전환해 보면 다음과 같다.

- 3.1에서 다룬 '데이터세트의 수량과 유효값 확인'의 경우 행정정보시스템 데이터베이스 설계 시 입력된 데이터가 무효한 상태가 되면 삭제를 하거나 상태 필드에 무효한 상태임을 명시하여 테이블 데이터만 보아도 유효한 정보를 구분할 수 있도록 한다. 또한, 필드에 유효한 범위의 값이 정해져 있는 경우에는 데이터베이스의 제약조건으로 설정하여 잘못된 값이 입력될 가능성을 배제해야 한다.
- 3.2에서 다룬 '일관된 코드 값 부여'의 경우, 하나의 데이터베이스 내에서, 혹은 하나의 행정정보시스템 내에서 여러 곳에 사용되는 동일 정보에 관한 코드 값은 통일된 값을 갖도록 해야 한다. 이의 효과적 실행을 위해, '성별'처럼 여러 곳에서 사용하는 참조 코드 값들을 한데 모은 통합코드 DB를 구축하고 데이터베이스 운용 시 이를 참조하여 실행하도록 한다.

www.kci.go.kr

- 3.3에서 다룬 '복합정보의 컴포넌트화'의 경우, 행정정보시스템의 데이터베이스 설계 시 지침이 필요하다. 즉, 하나의 칼럼이 다중 컴포넌트의 값을 조합한 복합정보로 설계하지 않도록 규칙을 명확히 해야 한다.
- 3.4에서 다룬 '날짜 데이터의 정밀도 결정'의 경우, 날짜 데이터를 생성하는 응용프로그램과 활용하는 응용프로그램 간에 정밀도에 관해 일관된 합의가 필요하다. 세기, 연, 월, 일, 시, 분, 초의 어느 단위까지 의미 있는 날짜 데이터를 입력하고 출력할 것인지를 정의하고 이를 관련된 모든 응용프로그램이 준수함으로써 데이터베이스에 정밀도가 다른 날짜 데이터가 발생하지 않도록 해주는 것이 필요하다.
- 3.5에서 다룬 '시점정보의 표준시간대 적용'의 경우, 행정정보시스템에서 만들어지는 모든 날짜 및 시점 데이터들은 반드시 시간대 정보를 함께 입력하여 저장하도록 해야 한다. 나아가 필요하다면 날짜 및 시점 데이터를 입력할 때 항상 표준시간대로 변환하여 저장하도록 할 수도 있다. 이를 통해 정확한 시점을 파악할 수 있도록 해준다.
- 3.6에서 다룬 '코드 값 설명정보'의 경우, 행정정보시스템의 데이터베이스 설계 시에 코드 값에 대해서는 표준화된 코드테이블을 생성하여 관리하는 것을 원칙으로 하여야 한다. 이 코드테이블에는 코드별 설명정보가 들어가게 된다. 만약 여러 기관이 공통으로 사용하는 코드값이라면 표준정부디렉토리시스템에서 관리해 주어야 하며, 각 기관의 행정정보시스템이 이러한 코드값 관리 기준을 잘 준수하는 것이 필요하다.
- 3.7에서 다룬 '메타데이터 확보'의 경우, 행정정보시스템이

www.kci.go.kr

나 데이터베이스가 자기 자신을 기술하는 메타정보를 충분히 보유하면서 운영되도록 해야 한다.

각 정부기관들은 데이터세트를 아카이브로 이관할 때 어떤 기준으로 품질을 체크하고 어떤 방향으로 보정할 것인지를 아카이브와 협의하여 지침으로 만들어 이행하여야 한다. 또한, 원활한 이관을 위해 행정정보시스템의 데이터베이스 품질요건을 강화하여 이관 시의 데이터 보정 및 품질 개선 노력을 절감할 필요가 있다.

행정정보 데이터세트를 추출하여 다른 시스템에 전송하는 요구사항은 단지 아카이브로의 이관에만 적용되는 것이 아니다. 공공 정보를 민간에게 제공하는 서비스가 활발해지면서 데이터세트 공공정보를 통합데이터베이스에 모으는 일에도 동일하게 적용될 수 있다. 미국과 영국 등 선진국에서는 미래지식정보사회 변화에 따라 국가가 보유한 공공 정보를 적극적으로 시민에게 개방하기 위한 정책을 수립하는 경향을 보이고 있으며, 우리나라도 이러한 추세를 따라가는 양상을 보이고 있다. 미국 예산관리국(OMB)은 오바마 대통령의 열린 정부 구현을 위해, 2009년 1월 '투명성 및 열린 정보에 관한 공람'을 발표하고 이를 추진하고 있다. 실행 지침으로는 부가가치가 높은 데이터세트를 선정 하고 연방데이터포털(Data.gov)에 등록하는 내용 등을 포함하고 있다. 이 사이트는 열린 정보 구현을 위해 효율성 높은 정부업무의 선진화 및 경쟁력 향상과 민주주의 발전에 기여하는 것을 목적으로 한다. 제공하는 원자료(raw data) 데이터 카탈로그, 도구 카탈로그, RSS 피드로 구분하여 정보를 제공하고 있다.[34] 영국 정부도 2010년 1월경 '범정부 공공 데이터 개방 · 공유 포

34) www.data.gov

www.kci.go.kr

털’을 공식 오픈하였다. 2009년 12월 발표한 ‘더 똑똑한 정보’가 되기 위한 실천 전략의 하나로 미국의 data.gov를 벤치마킹 모델로 선정하여 추진되었다. 현재 구축된 포털에서는 원하는 정보와 데이터를 쉽고 편리하게 검색 및 활용이 가능한 상태이다[35]. 우리나라도 행정안전부가 2010년 3월 『공공정보 민간활용 촉진 종합계획』을 문화체육관광부 · 방송통신위원회와 공동으로 발표하였다. 이 계획에는 공공정보에 대한 접근성 강화, 공공정보의 제공과 활용을 위한 제도 정비, 공공정보의 품질제고, 그리고 민간 활용을 지원하는 내용이 담겨 있다.[36]

공공정보를 민간에 제공한다는 초점에서의 데이터세트의 대상, 제공방식은 기록관리 측면에서의 대상 및 방식과 차이점이 있으나 데이터세트를 추출하여 집적하고 제공하는 과정의 구조적, 기술적 측면에서는 유사점을 가진다. 특히, ETT과정에서 데이터의 보정 및 품질 개선이 필요할 것이라는 점은 동일한 요구사항이 될 것으로 예측된다. 앞으로 민간에 적극적으로 정보를 제공하겠다는 정부의 방침이 전반적인 업무 방식, 정보 및 기록을 형성하는 인프라에 큰 영향을 미칠 것으로 예상된다. 이러한 흐름을 간과하지 말고 기록관리적인 측면에서 기여할 수 있는 사항을 예측하고 그동안 축적된 기술 및 경험치를 공유할 수 있는 노력이 필요할 것으로 보인다. 이 논문에서 제시한 데이터세트의 보정 및 품질 개선 사례는 향후 데이터세트 아카이브 구축뿐 아니라 공공정보의 공유활성화를 위한 기법으로 유용성을 발휘하게 될 것이다.

35) www.data.go.uk 참고
36) 한겨레신문, 2010년 월 9일자 “버스시간·휘발유값 등 공공정보 민간에 개방”

www.kci.go.kr

ABSTRACT

A Study on Data Adjustment and Quality Enhancement Method for Public Administrative Dataset Records in the Transfer Process - Based on the Experiences of Datawarehouses' ETT

Yim, Jin-Hee· Cho, Eun-Hee

As it grows more heavily reliant on information system, researchers seek for various ways to manage and utilize of dataset records which is accumulated in public information system. It might be needed to adjust date and enhance the quality of public administrative dataset records during transferring to archive system or sharing server. The purpose of this paper is presenting data adjustment and quality enhancement methods for public administrative dataset records, and it refers to ETT procedure and method of construction of datawarehouses.

It suggests seven typical examples and processing method of data adjustment and quality enhancement, which are (1) verification of quantity and data domain (2) code conversion for a consistent code value (3) making component with combinded information (4) making a decision of precision of date data (5) standardization of data (6) comment information about code value (7) capturing of metadata. It should be reviewed during dataset record transfer.

This paper made Data adjustment and quality enhancement requirements

www.kci.go.kr

for dataset record transfer, and it could be used as data quality requirement of administrative information system which produces dataset.

Key words: Dataset, eletronic record transfer, Data Adjustment and Quality Enhancement, Datawarehouse, ETT

www.kci.go.kr

www.kci.go.kr