韓國컴퓨터情報學會 論文誌
第16卷 第6號, 2011. 6.

2011-16-6-3-1

데이터 품질을 고려한 국가R&D정보 데이터베이스의 통합 사례 연구– NTIS 데이터베이스 통합 사례

신 성 호*, 윤 영 준**, 양 명 석**, 김 진 만**, 손 강 렬**

A Data Cleansing Strategy for Improving Data Quality of National R&D Information - Case Study of NTIS

Sung-Ho Shin*, Young-Jun Yoon**, Myung-Suk Yang**, Jin-Man Kim**, Kang-Ryul Shon**

요 약

데이터 품질관리 관점에서 볼 때, 데이터의 품질은 품질정책, 품질조직, 업무프로세스, 업무규칙 등 여러 요인에 의해 영향을 받는다. 이중에서도 업무규칙은 실제 데이터를 조작하는 행위의 지침이 되는 것으로써 데이터 품질에 직접적인 영향을 미친다. 여러 기관의 데이터베이스를 통합하여 단일의 데이터베이스를 구축하는 경우에는 더 신중하게 업무규칙을 수립할 필요가 있다. 분산된 데이터베이스 내에 있는 데이터를 단일의 데이터베이스로 통합한다는 것은 단순히 데이터의 통합만을 의미하는 것이 아니라 상이한 스키마, 코드 체계, 데이터 표준 등을 사전에 고려해야 함을 의미한다. 이런 요소들을 고려하더라도 데이터 자체는 형식, 단위, 표현 등에 따라서 다양한 모습을 가진다. 결국 데이터베이스의 구조적인 문제와 데이터 자체의 의미적인 문제가 데이터베이스 통합과 통합된 데이터베이스 내 데이터의 품질 제고를 위한 선결 과제라 할 수 있다. 이러한 문제들을 해결하기 위해서는 먼저 통합 시 통합 모델의 설계가 필요하고, 통합된 데이터베이스의 데이터에 대한 정제가 필요하다. 범부처적으로 분산되어 있는 국가R&D정보를 수집하여 서비스하는 국가과학기술종합정보서비스(NTIS)도 여러 기관에 존재하는 데이터베이스를 연계·통합하여 단일의 데이터베이스를 구축하였다. NTIS의 사례를 통해 체계적인 통합 모델 수립과 정제에 의해 통합된 데이터베이스의 데이터는 그렇지 않은 데이터보다 정확도 측면에서 품질이 제고되었음이 입증되었다.

▸ Keyword : 데이터품질, 데이터정제, 데이터베이스통합

Abstract

On the point of data quality management, data quality is influenced by quality policy, quality organization, business process, and business rule. Business rules, guide of data manipulation,

• 제1저자 : 신성호 • 교신저자 : 손강렬
• 투고일 : 2011. 02. 14, 심사일 : 2011. 02. 22, 게재확정일 : 2011. 03. 03.
* 한국과학기술정보연구원 정보기술연구실(Dept. of Information Technology Research, KISTI)
** 한국과학기술정보연구원 NTIS사업단(Dept. NTIS, KISTI)

have effects on data quality directly. In case of building an integration database among distributed databases, defining business rule is more important because data integration needs to consider heterogeneous structure, code, and data standardization. Also data value has various figures depended on data type, unit, and transcription. Finally, database structure and data value problem have to be solved to improve data quality. For handling them, it is needed to draw database integration model and cleanse data in integrated database. NTIS(stands for National science and Technology Information Service) has an aim to serve users who need all information about national R&D by internet, and for that aim, it has a integrated database which has been made with several database sources. We prove that database integration model and data cleansing are needed to build a successful integrated database through NTIS case study.

▸ Keyword : Data Quality, Data Cleansing, Database Integration

I. 서 론

데이터 정제는 데이터 품질을 높이기 위해 데이터 오류나 불일치를 발견하고, 처리하는 것이다. 데이터 입력 과정에서의 오타, 손실 정보, 오류 데이터들은 항상 발생할 수 있기 때문에 데이터 품질 문제는 파일이나 데이터베이스와 같은 단일의 데이터 집합에서 늘 발생되는 문제이다. 데이터웨어하우스, 분산·통합된 데이터베이스, 글로벌 웹 기반 정보시스템 등과 같이 여러 개의 데이터 소스가 통합되어야 하는 경우가 발생한다면, 데이터 정제가 더욱 중요한 문제가 될 수 있다. 데이터 소스들이 통합되면 중복된 데이터들이 존재할 수 있기 때문이다. 정확하고 일관된 데이터를 제공하기 위해, 다른 형태의 데이터 통합과 중복 데이터의 제거와 같은 작업들이 필요하다.

데이터 정제는 여러 가지 요구사항들을 충족시켜야 한다. 무엇보다 개별 데이터 소스들 간 존재하거나 통합 과정에서 발생할 수 있는 데이터 오류들과 불일치를 찾아내어 제거해야 한다. 또한 수작업을 최소화하기 위해 툴의 지원을 받아야 하고, 추가되는 데이터 소스들을 쉽게 반영할 수 있어야 한다. 더 나아가서, 스키마와 연계된 데이터 변환과 함께 이루어져야 한다.

본 연구에서는 정부 부처·청에 분산되어 존재하고 있는 국가R&D정보 데이터베이스의 통합 사례를 통해, 첫째, 데이터 정제과정을 접목시킨 통합 데이터베이스 구축 모델을 제시하고, 둘째, 이 통합 모델을 통한 데이터베이스 통합 결과, 데이터 품질이 향상되었음을 증명하고자 한다. 셋째, 이러한 결과들을 통해 데이터베이스 통합 모델 개발 및 데이터 품질 향상을 위한 데이터 정제의 중요성에 대해 고찰하고자 한다.

II. 데이터 정제 및 데이터 품질

1. 데이터 정제 개요

데이터 정제는 필요에 따라 그 목적이 다르기 때문에 여러 가지 정의가 가능하다. 본문에서는 데이터 정제에 대한 일반적인 정의와 정보시스템 유형에 따른 정제 개념에 대해 살펴본다.

Levitin and Redman[1]의 연구에서는 데이터 정제를 '전사적 데이터 품질관점에서 데이터 생성주기를 데이터 획득과 사용으로 구분하고, 평가, 분석, 조정, 제거 등 일련의 연속된 과정'으로 정의하였다. Simoudis et al.[2]의 연구에서는 데이터 정제를 '데이터베이스 검사, 오류와 누락 데이터 검색 그리고 수정하는 과정'으로 보았다. Guyon et al.[3]의 연구에서는 데이터 정제를 '기계학습 기술을 이용해 의미가 없거나 잘못된 데이터 유형을 발견하고 분류하는 작업 과정'으로 이해하였다. Kimball[4]의 연구에서는 데이터 정제를 ' 세분화(segmentation), 표준화(standardizing), 검증(verifying), 일치(Matching), 군집(Clustering)의 6단계를 거쳐 오류를 제거하는 과정'으로 정의하였다. Hernandez and Stolfo[5]의 연구에서는 데이터 정제를 '데이터 병합과 제거, 그리고 중복 가능성이 있는 레코드 정렬을 통해 이를 인접한 레코드로 만들고, 이 중 품질이 낮은 레코드를 제거하는 것'으로 정의하였다. Galhardas et al.[6]의 연구에서는 데이터 정제를 '오류와 불일치 데이터 제거와 필드 내 문제를 해결하는 과정'으로 정의하였다. Rahm and Do[7]의 연구에서는 데이터 정제를 '데이터 품질을 높이기 위해 데이터 오류나 불일치를 발견하고, 처리하는 것'으로 정의하였다. Chapman[8]의 연구에서는 데이터 정제를 '부정확, 불완전, 또는 비논리적인 데이터를 결정하여 발견된 오류와 누락된 부

분의 수정을 통해 품질을 개선하는 일련의 과정'으로 해석하였다.

위 학자들의 견해를 종합해 보면, '데이터 정제란 데이터 오류나 불일치를 발견하고, 처리하는 것'이라고 정의할 수 있는데, 일반적으로 데이터 정제에는 데이터 품질을 높이기 위한 노력의 개념이 포함되어 있다. 따라서 부정확, 불완전, 또는 비논리적인 데이터를 결정하여 발견된 오류와 누락된 부분의 수정을 통해 품질을 개선하는 일련의 과정으로 정의된다.

2. 데이터 정제 단계

데이터 품질을 보장하기 위하여 이루어지고 있는 데이터 정제는 일반적으로 다음과 같이 크게 세 가지 과정으로 살펴볼 수 있다[9][10].

1단계. 데이터 분석(data analysis)

데이터 정제 과정의 우선적인 작업은 정제하고자 하는 데이터 속에 어떤 유형의 오류가 있는지를 파악하는 일이다. 정제할 데이터의 유형을 파악하여 작업의 목표 설정 및 작업에 필요한 참조 데이터를 파악하는 등 전체 작업을 준비하는 단계이다.

2단계. 데이터 세분화(data segmentation)

세분화 과정은 일관된 형태가 아닌 자유로운 형태로 입력한 데이터값들을 정제할 수 있는 작업 단위로 분할하는 과정이다. 다양한 표현으로 입력된 원천 데이터를 변환 가능한 단위까지 분할하는 과정이 선행되어야 한다.

3단계. 데이터 표준화(data standardization)

데이터 표준화란 세분화된 각 데이터 값에서 글자 사이의 모든 공란을 제거하고 의미적으로는 같으나 여러 가지 표기방식으로 표현된 데이터를 통일된 표기로 바꾸어 일치시키는 과정이다. 아파트, APT, apt, @ 등 다양한 형태로 표기된 값을 통일된 표기로 변환하는 작업이 이 과정에 포함된다.

4단계. 오류 검출과 수정(error detection and correction)

오류 검출과 수정 단계는 표준화단계까지 거친 데이터 속에 포함되어 있는 오류를 검출하고 수정하는 단계이다. 오류의 유형은 특이값(outlier), 불일치값(inconsistency values), 필드에 값이 없는 것(missing values) 등이다.

5단계. 무결성 체크(integrity checking)

무결성은 개체, 도메인, 참조, 사용자 정의 무결성으로 구분하여 정의한다. 속성 간에 일치되지 않는 값이 있는지 그리고 같은 개체를 표현하는 여러 개의 중복 레코드가 있는지를 검사하는 단계이다.

데이터 정제의 궁극적인 목적은 데이터의 품질을 높이는데 있다. 데이터 품질에 대한 일반적인 정의를 내리기는 어렵다. 데이터 품질에 관한 기존 연구는 정확성이란 개념 하에 정의를 내리고 있으나, 최근 연구는 데이터가 사용되는 상황에 따라 품질이 달라질 수 있다는 상황론적 시각을 제시하고 있다. 즉, 절대적인 관점에서는 오류 데이터가 아닐지라도, 특정 조직이나 서비스 중심의 관점에서 보면, 업무규칙에 어긋난 데이터이기 때문에 오류 데이터가 되기도 한다. 따라서 데이터 정제도 정확성에 기초한 정제와 더불어, 조직이나 서비스에서 요구하는 업무규칙에 기초한 사용 적합성 위주의 정제가 필요하다.

III. 데이터베이스 통합 및 관련 개념

1. 데이터베이스 통합 및 방법

데이터웨어하우스 구축, 글로벌 웹 기반 정보시스템 구축 등과 같은 작업을 위해 여러 개의 데이터 소스가 통합되어야 하는 경우가 발생한다. 서로 다른 데이터 소스들은 각각의 필요에 의해 독립적으로 개발되고, 유지되어 왔기 때문에 데이터관리시스템, 데이터 모델, 스키마 설계, 그리고 실 데이터는 매우 이질적인 모습을 보이게 된다. 데이터베이스 통합에서 분산된 데이터베이스들을 한데 모으기 위해서는 스키마의 통합과 더불어 데이터의 정제가 더욱 중요한 문제가 될 수 있다. 여러 데이터 소스들이 통합되면 중복된 데이터들과 데이터 오류들이 존재할 수 있기 때문이다. 정확하고 일관된 데이터를 제공하기 위해, 다른 형태의 데이터 통합과 중복 및 오류 데이터의 제거와 같은 작업들이 필요하다.

지금까지 이질적인 데이터베이스 간 통합을 위한 많은 방법들이 제안되었다. 정보시스템 통합방법은 크게 3가지로 분류할 수 있다[11]. 첫째, 고전적 접근 방식이다. 이 방식은 다양한 지역 소스(local source) 스키마의 차이를 해결하기 위한 하나의 전역 스키마(global schema), 즉, 지역 소스들을 논리적으로 통합하여 하나의 일관성 있는 통합 스키마의 작성이 핵심이 된다. 둘째, 고전적 접근방식이 유발했던 데이터베이스 사용자 이질성의 문제를 해결한 연합된 접근법이 있다. 하지만, 여전히 전역 스키마의 방식을 사용하고 있고, 전역 스키마 또한 여전히 정적이기 때문에 지역 소스의 추가나 수정 시 전역 스키마의 갱신이 요구되어진다. 셋째, 분산 객체관리방식이 있다. 이는 분산되고 이질적인 데이터베이스를 분산 객체 공간의 객체 집합 모델 기반으로서, 연합된 접근

방법을 일반화시켰다는데 의의가 있다.

공통적으로 제시되는 개념으로써, 데이터베이스 통합을 위해서는 스키마통합이 선행되어야 한다[8]. 스키마 통합은 서로 이질적인 데이터베이스들의 스키마 구조를 통일시켜 하나의 일관된 데이터베이스로 만드는 것을 의미한다.

스키마 통합과정에서 주요한 문제들은 이름 충돌과 구조적인 충돌이다. 이름 충돌은 의미는 다르지만, 동일한 속성이름으로 각각 사용되는 경우(동음이의어)와 속성이름은 다르지만 같은 의미를 가지는 경우(이음동의어)이다. 구조적인 충돌은 매우 많은 변수로 인해 발생되며, 동일한 객체지만 다른 소스들에서 다른 모습으로 존재하기 때문에 발생한다. 다른 콤포넌트 구조, 다른 데이터 타입, 다른 무결성 제약조건, 코드값 등이 그 예이다[9].

2. 데이터베이스 통합 시 발생할 수 있는 특성

데이터베이스 통합이라는 개념 자체가 여러 개의 데이터 소스와 관련되어 있다는 사실을 내포하고 있다. 때문에 시스템 간 데이터를 통합할 때 발생할 수 있는 이질적인 특성이 기본적으로 존재한다[12].

본 연구에서는 시스템적 환경의 이질성 보다는 데이터 관점에서 발생 가능한 이질적인 특성을 주로 다루었다.

○ 데이터 구조의 이질성

데이터 구조의 이질성은 데이터 관리에 대한 표준화 방안을 제시하지 못하기 때문에 발생한다. 독립적인 내부 표준에 의해 설계된 시스템의 경우 체계적이고 효율적인 데이터 관리는 가능하지만 같은 분야의 시스템 간의 데이터 교환 및 통합에는 문제점을 갖고 있다.

서로 다른 두 시스템에서 유사한 데이터를 관리하더라도 업무 및 데이터베이스 설계자에 의해 테이블 및 필드의 특성은 다양하게 설계되어 테이블과 필드의 구조, 테이블간의 관계 및 주키와 외래키 설정은 서로 다르게 설정된다.

○ 데이터 정의의 이질성

데이터 정의의 이질성은 데이터를 위한 데이터 즉, 메타데이터의 의미가 서로 다르게 정의된 것으로서, 데이터를 주고받는 사람들이 서로 다른 이름으로 데이터를 정의할 경우 동일한 데이터를 같은 의미로 해석할 수 없는 문제가 발생한다.

○ 데이터 표현의 이질성

메타데이터는 동일하지만 데이터의 형식 및 단위 등이 서로 다르게 정의되어 동일한 데이터가 서로 다르게 표현될 수 있다. 상호간에 데이터를 교환 및 통합해야하는 경우 이러한 데이터의 형식 및 값의 불일치는 데이터 간의 호환성 유지를 어렵게 한다. <표1>은 데이터 표현의 이질성에 대한 문제를 세 가지로 분류한 것이다.

표 1. 데이터 표현의 이질성 분류
Table 1. Heterogenous classification

분류	설명
데이터 형식의 이질성	날짜의 경우 데이터의 형(Type)이 Date로 설계되어 사용되는 경우와 일반적으로 문자열(String)로 설계되어 사용되어질 수 있다.
데이터 단위의 이질성	데이터의 값이 수치 또는 양을 나타내는 경우 단위의 이질성이 발생할 수 있다. 예를 들면, 진료비의 경우 만원이라는 값을 "10000원" 또는 "10,000" 또는 "10$"로 표현할 수 있다.
데이터 표현의 이질성	예를 들면, 약물검사결과에서 동일한 데이터를 "양성" 또는 "positive" 또는 "+"로 표현할 수 있다.

이러한 데이터 이질성 문제를 해결하기 위해서는 시스템 및 데이터 측면에서 표준 제정 및 적용이 활발히 이루어져야 한다. ISO/IEC에서는 정보를 주고받는 사람들이 동일한 정보를 같은 의미로 정보를 해석할 수 있도록 데이터를 형식화하는데 사용되는 데이터 요소의 서술 및 형식화, 유지보수에 대한 지침을 분류하여 제공한다. <표2>에서는 관련된 표준 중 ISO/IEC 11179에 대해 정리하였다.

표 2. ISO/IEC 11179 내용
Table 2. Content of ISO/IEC 11179

표준번호	내용
11179-1	데이터 요소의 생성과 표준화를 위한 프레임워크
11179-2	도메인 식별을 위한 개념의 분류
11179-3	기본 데이터 요소 속성
11179-4	데이터 요소의 형식화를 위한 규칙과 지침
11179-5	데이터 요소를 위한 명명과 식별
11179-6	데이터 요소의 등록

Ⅳ. 국가R&D정보 데이터베이스 통합 및 정제 사례

1. 국가과학기술종합정보서비스(NTIS)

교육과학기술부와 한국과학기술정보연구원은 2006년부터 국가과학기술종합정보서비스(National science and Technology Information Service, 이하 NTIS)를 구축하여 서비스 중에 있다.

NTIS의 서비스 대상은 국가R&D사업 관련 메타정보인

데, 국가R&D사업 관련 정보란 범부처적으로 추진되고 있는 R&D사업 또는 과제들의 수행 전후로 발생되는 과제정보, 성과정보, 인력정보, 장비정보 등을 의미한다. 이렇게 부처별·기관별로 개별 관리되고 있는 국가R&D 관련 정보를 공유·공동 활용함으로써 국가R&D투자 효율성을 제고시킬 수 있다. <그림1>은 국가R&D정보 지식포털(NTIS) 개념도로서 국가연구개발 사업 및 과제가 어떻게 수집되어 활용 및 서비스 되는지를 보여주고 있다[13].

그림 1. 국가R&D정보 지식포털(NTIS) 개념도
Fig. 1. Concept Diagram of NTIS

2. 국내·외 현황

국내에서 국가R&D정보 데이터베이스의 통합 사례를 찾아보기는 어렵다. 한국연구재단, 중소기업청, 기술표준원 등에서 각 기관의 특성에 맞는 국가R&D정보를 선별적으로 구축하고 있다(표3).

표 6. 국가R&D정보 데이터베이스 구축 현황
Table 3. The Cases of National R&D Database

분야	구축기관	비고
R&D인력정보	한국연구재단, 한국산업기술평가관리원, 정보통신연구진흥원, 한국보건산업진흥원 등	NTIS 연계
장비·기자재 정보	한국기초과학지원연구원	NTIS 연계
	기술표준원, 한국산업기술진흥원, 중소기업청 등	-
사업관리 정보	한국과학기술기획평가원 등	NTIS 연계
성과정보	한국과학기술정보연구원, 정보통신연구진흥원 등	NTIS 연계
	국가생물자원정보관리센터, 생물자원센터 등	-
공통기반	한국정보사회진흥원, 기술표준원, 한국정보통신기술협회 등	-

해외에서도 NTIS와 같이 국가R&D정보를 범부처적으로 통합한 사례는 찾아보기 어렵다. 일본 문부과학성의 e-RAD, 미국 과학재단의 RaDiUS와 NTIS, 유럽연합의 CORDIS와 DRIS 등이 존재하지만, 국가R&D 데이터베이스를 통합하기 위한 목적보다, 개별적으로 국가R&D 프로젝트를 관리하기 위한 목적으로 개발 및 운영되고 있다. 또한 일부는 많은 데이터베이스를 통합하고 있지만, 현행 연구 프로젝트들의 연구정보를 제공하기 위한 목적으로 존재한다.

3. 국가R&D정보의 특성

국가R&D정보란 국가 차원에서 예산 투입이 된 연구개발사업의 메타정보로써 연구개발사업의 시작부터 끝까지 발생되는 모든 정보들을 의미한다. 국가R&D정보는 다음과 같은 특성을 가진다. 첫째, 국가R&D정보는 그 정보가 방대하여 정확하게 한정짓기가 쉽지 않다. 국가R&D사업의 방향에 따라 국가R&D정보의 범위 및 유형이 결정되므로, 관련 시스템과 스키마는 언제든지 확장가능 해야 한다. 둘째, 국가R&D사업을 수행하는 정부 부처·청에 분산되어 존재한다. 이로 인하여 각 부처·청별로 관리하는 정보 항목이 다르고, 스키마/코드도 다르다. 즉, 스키마, 표준, DBMS, 어플리케이션 등 데이터 형태 및 관련 시스템이 매우 이질적이다. 셋째, 국가R&D사업을 수행하는 정부 부처·청에 분산되어 존재하기 때문에, 데이터베이스 접근 통제 및 지역 자치성을 보장해 주어야 하는 어려움이 있다. 즉, 데이터베이스의 통합을 위해서는, 필요한 경우, 지역 스키마의 변경이 필요하고, 데이터베이스의 접근도 필요하지만, 관련 기관들의 협조가 어려운 경우도 발생한다. 따라서 소스 데이터베이스를 가지고 있는 각 기관의 데이터베이스에 접근 및 스키마의 변경 없이 데이터베이스 통합을 이루어내야 한다는 어려움이 있다.

기업정보의 통합과 국가R&D정보의 통합 간 가장 큰 차이가 이 부분이다. 기업정보의 통합은 통합 대상이 되는 데이터 소스가 동일한 조직 내에 있거나, 다른 조직에 존재하더라도 데이터베이스에 접근 및 스키마의 변경이 가능하다. 때문에, 데이터베이스의 통합을 위해 목적에 따라서 웹 기반 사용자 인터페이스, 응용 수준 통합, SQL 미들웨어를 통한 통합, 지역 통합을 통한 뷰 제공, 공유 저장소의 방법 등 다양한 데이터베이스 통합 기법들을 적용할 수 있다[14]. 하지만, 국가R&D정보 데이터베이스의 통합을 위해서는 이와 같은 시스템 및 어플리케이션에 의한 자동화된 방법보다는 고유의 통합 모델을 설계하여 통합을 진행하는 수동적인 통합이 더욱 적합하다.

4. 국가R&D정보 데이터베이스 통합 모델 설계

NTIS 통합 데이터베이스 구축을 위한 기본 과정은 데이터 수집, 데이터 정제, 서비스DB 구축 세 단계로 이루어진다.(그림2)

국가R&D정보의 특성에서 오는 문제점들을 해결하고 공동 활용을 위해서는 스키마와 코드에 대한 표준이 선행되어야 하고, 고유의 데이터베이스의 통합 모델의 설계가 필요하다. NTIS에서는 부처 간 공동 활용을 위한 정보항목을 마련하고자 NTIS 세부시스템을 통해 서비스 될 항목을 중심으로 범부처 국가R&D정보 340개 항목을 확정하였다(2010년 현재). 범부처 국가R&D정보 항목은 과제, 성과, 인력, 장비·기자재 등의 분야에서 기존 조사분석 정보항목을 모두 포함한 총 340개 항목으로 각 부처·청을 통해 수집하고 있다[15].

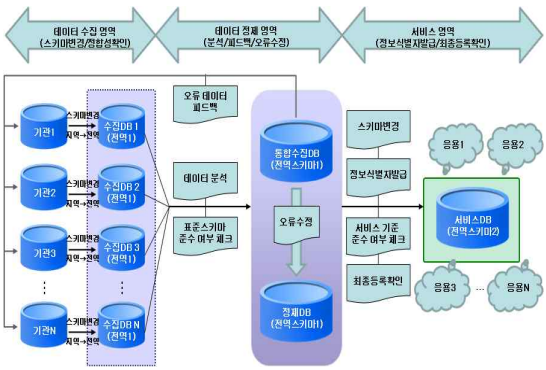

그림 2. 국가R&D정보 데이터베이스 통합 모델
Figure 2. Model of database integration for R&D data

○ 데이터 수집 영역

NTIS에서는 국가R&D 데이터의 효율적인 수집을 위해 각 부처·청별로 국가R&D 데이터의 관리기관(대표전문기관)을 두고, 해당 부처·청별로 대표전문기관을 통해 NTIS로 데이터를 제공하도록 하였다. 분산되어 있는 대표전문기관DB로부터 데이터를 수집하기 위해 범부처 국가R&D정보의 표준화, 정보연계방식 및 절차 표준화, 정보연계S/W 구축 등을 수행한다. 즉, 각 대표전문기관의 Legacy DB에 존재하는 데이터를 NTIS 스키마(전역1스키마)로 변환하고, 코드성 데이터는 NTIS 표준코드로 변환하여 NTIS 정보연계를 통해 NTIS 수집 DB로 전송한다.

○ 데이터 정제 영역

수집된 데이터는 서비스DB로 구축되기 전 정제 과정을 거치는데, 정제 과정 전에 수집된 데이터에 대한 분석 및 확인 과정을 거친다. 수집된 데이터가 NTIS 스키마 및 코드에 맞게 잘 변환되었는지 일종의 검수 과정을 거친다. 데이터 검수를 위해서는 데이터 분석이 선행되어야 하는데, 스키마 변환, 코드 변환, 필수항목 수집 여부 확인, 가베지 데이터 확인 등의 분석을 수행한다.

이 과정에서 오류 데이터로 판별된 데이터는 각 해당 기관으로 통보되어 데이터 보완 요청이 이루어진다. 이 과정이 데이터 피드백 과정이다. 기본적으로 데이터 발생 및 1차 수집기관이 정보연계기관이므로 1차적인 정제가 매우 중요하며, 데이터 내용의 수정 및 발생은 정보연계 기관에서 수행하는 것이 가장 정확하다. 1차 정제 및 데이터 피드백을 통해 서비스 데이터로 준비된 데이터는 2차 정제 과정(내부 정제)을 거치게 된다. 이 과정에서는 데이터 보완과 수작업 검증이 이루어진다. 데이터 검증이란 정제 작업자가 기관으로부터 수집된 데이터를 NTIS 데이터정제시스템을 통해 건건이 확인하여 데이터를 확인하는 과정이다.

이러한 과정들을 선행연구에서 제시한 데이터 정제 단계와 비교하면 순서가 조금 다를 수 있지만, 각 단계에서 필요한 작업들이 모두 수행이 되고 있다. 데이터 분석은 NTIS에서 데이터 수집 발생되는 작업으로써, 데이터 정제 작업을 위한 선행 작업이다. NTIS에서 데이터 세분화 작업은 참여연구원 주소나 소속기관명 항목에서 주로 발생된다. 참여연구원 주소는 스키마에서 텍스트(Char, Varchar2)로 수집되도록 되어 있기 때문에 주소1과 주소2를 구분하지 않고, 주소1에 주소2까지 모두 작성하는 경우가 발생한다. 이 경우 주소1과 주소2를 나누는 작업이 필요하다. 소속기관명도 소속기관명 항목에 기입되어 있지 않고, 주소에 기입되어 있는 경우가 있기 때문에 데이터 세분화 작업을 해 주어야 한다. 데이터 표준화를 위해서 사전에 표준 스키마 및 코드를 정의하여 1차 수집 기관인 대표전문기관에 배포하고 있다. 오류 검출과 수정은 기계정제와 수작업정제를 통해 수행되고 있다. 마지막으로 무결성 체크는 데이터 수집 영역, 데이터 정제 영역, 서비스 영역 각각에서 필요에 의해 수시로 이루어지고 있다.

○ 서비스 영역

서비스 영역에서는 스키마 변환(전역스키마1→전역스키마2), 정보식별자 부여, 서비스 기준 준수 여부 체크, 최종등록확인 작업을 수행하고, 최종등록확인이 된 데이터에 한해서 서비스DB로 구축된다. 전역스키마1은 데이터 수집을 용이하게 하기 위해 서비스를 고려하지 않은 스키마이므로, 서비스를 위해서는 서비스를 고려한 스키마로 변경이 필요하다. 정보식별자는 과제, 성과, 인력 데이터가 상호 참조가 되어 연계가 되도록 하기 위해 덧붙여지는 관리성 데이터 항목이다.

본 모형의 특징은 데이터베이스 통합의 최종 목표를 고품

질의 데이터 확보에 초점이 맞추어져 있다. 데이터의 품질을 높이기 위해 데이터베이스 통합 시 필수 작업들을 정의하고, 이 작업을 수행하기 위해 작업을 여러 구간으로 나누어 놓았다. 각 구간에는 해당 작업의 결과물을 저장하는 데이터베이스가 존재한다. 자원 및 작업의 효율성만을 생각하면, 구간의 구분 없이 한 구간에서 모든 작업을 수행할 수도 있지만, 이렇게 될 경우, 국가R&D정보의 특성 상 많은 데이터가 수집되기 어렵게 된다. 국가R&D정보는 각 부처 및 기관별로 분산되어 있고, 스키마 및 표준도 각각 다르기 때문에 수집되는 첫 단계부터 모든 서비스 요구사항을 적용시킨다면, 이 요구조건을 충족하는 데이터가 많이 부족해질 것이다. 따라서 작업 구역을 여러 개로 나누어 놓음으로써 가능한 한 많은 데이터가 수집되도록 하였다. 수집된 데이터는 최소한의 요구조건만 충족시킨 상태이기 때문에, 표준스키마 및 표준코드에 위배되는 데이터들이 여전히 존재한다. 이러한 오류 데이터들은 데이터를 제공한 기관으로 다시 보내져서 재 수집되고, 일부 간단한 오류들은 미리 정의해 둔 정제 규칙에 의해 데이터를 정제(가공)를 한다. 정제된 데이터들에 대해서 서비스 기준 준수 여부를 체크한 후, 기준을 충족시킨 데이터만 서비스DB로 구축되게 된다.

기존의 데이터베이스통합과 관련된 연구들은 데이터의 품질 측면보다는, 효율성을 중점으로 자동화된 데이터베이스통합에 초점이 맞추어져 있었기 때문에 통합된 데이터베이스의 데이터 품질을 높이기 위해서는 통합 이후 별도의 정제 과정이 필요하다. 이 경우는 서로 다른 두 조직 간의 데이터베이스의 통합이라든지, 한 개의 조직 내에서도 부서의 통합으로 인한 데이터베이스 통합의 경우처럼 최초 1회성이 많기 때문에, 선통합·후정제도 가능하다. 하지만, NTIS와 같이 지속적으로 이질적인 데이터베이스 데이터들의 연계·통합하여 실시간으로 서비스 해야 되는 경우에는 데이터베이스의 통합과 동시에 데이터의 품질을 고려한 데이터베이스 통합 모델의 설계가 필요하다.

V. 데이터베이스 통합 모델 검증

1. 데이터 오류 유형 및 정제 규칙 정의

데이터 오류를 결정짓는 요소는 데이터 내용, 데이터 포맷, 표준 코드값, Null값 여부, 테이블 간 연계성(PK, FK) 등이 있다. 기준 데이터와 내용적인 측면에서 일치해야 되고, 데이터 형태가 동일해야 되며, 코드성 데이터는 최종 전송 대상 데이터베이스를 사용하는 응용에서 미리 정의해 놓은 코드 값들과 일치해야 한다. 테이블 간 연계가 되지 않은 경우도 오류라 볼 수 있다. 테이블 간 연계성은 연계되는 테이블 간 주키(PK: primary key)와 참조키(FK: foreign key)의 일치 여부로 판단할 수 있다. 이렇듯 데이터베이스 구축의 궁극적인 목표가 되는 데이터 품질 향상을 위해서는 정확한 데이터 입력도 중요하지만, 입력된 데이터의 오류 부분들을 체계적으로 정제하는 것도 중요하다.

고품질의 NTIS 통합데이터베이스 구축을 위해 시스템에 의한 정제 뿐 아니라 정제작업자에 의한 수작업 정제도 필요하다. 시스템에 의한 정제는 정형화된 틀 속에서의 정제만 가능하지만, 비정형화된 데이터를 정제하는 데는 한계가 있다. 따라서 시간과 비용이 들지만 고품질의 데이터를 확보하기 위해 정제작업자에 의한 수작업 정제도 필요하다.

시스템 정제와 정제작업자에 의한 수작업 정제 모두 일관된 정제 규칙이 필요하다. 각 기관들로부터 수집된 데이터를 분석해보면, 다양한 형태의 오류 유형을 발견할 수 있다. 정제 규칙에는 '특정 데이터가 오류이다 아니다에 대한 기준, 오류이면 어떤 유형의 오류인지, 그리고 각 유형별 오류를 어떻게 정정할 것인가'에 대한 세부적인 지침까지 포함하여야 한다.

데이터 정제 규칙을 정의하기 위해서는 발생 가능한 오류 유형에 대한 예측이 필요하다. 데이터베이스 통합 시 발생 가능한 오류 유형은 <표4>과 같다.

표 4. 데이터 오류 유형(수집/정제 영역)
Table 4. Types of Data error
(area of data gathering & cleansing)

구분		오류 유형
내용 오류 (유형1)		Not Null 항목이나 NULL 값
		Not Null 항목은 아니지만 NULL 값
		Garbage 성 데이터
스키마 오류 (유형2)		테이블 간 연결키값 없음
		중복 레코드
코드 오류 (유형3)		코드 항목이지만 명칭으로만 존재
		코드항목이지만 NTIS표준코드로 매핑 불가
형식 오류 (유형4)	날짜	YYYY, YYYYMM, YYYYMMDD 형식 오류
	숫자	정수 단위 오류
		소수점 이하 자릿수 오류
		숫자가 아닌 데이터
	문자	특정문자 반복(Garbage성으로 판단)
		특정문장 반복(Garbage성으로 판단)
		유니코드, 특수문자, CR-LF 포함
기타오류 (유형5)		(5-1) 논문, 특허 저자의 구분자(;) 오류
		(5-2) 사업부처명 및 사업부처 코드 오류
		5-2-1.사업부처명만 수집
		5-2-2.사업부처코드만 수집
		5-2-3.사업부처명과 코드가 각각 다름

NTIS 데이터는 수집 주체와 서비스 주체가 다르다는 특성이 있다. 따라서 서비스 주체 쪽에서는 오류 데이터에 대한 정제가 매우 제한적일 수밖에 없고, 오류 유무를 판단하기도 어렵다. 특히 데이터의 내용 오류, 스키마 오류, 코드 오류에 대해서 수정이 불가하다. 이런 오류들이 발생하면 데이터를 재 수집하는 것이 원칙이다. 형식 오류의 경우만 일부 정제가 가능하다.

데이터 정제는 오류 유형 각각에 대해 이루어져야 한다. 각 오류 유형에 대한 정제 규칙은 우선적으로 NTIS 표준 스키마 및 표준 코드에 따라 정제하는 것이 기본원칙이다. 정제 시 이 기준에 부적합한 데이터에 대해서는 추가적인 규칙을 정의하였다. <표5>에서는 NTIS에서 데이터베이스 통합 시 적용된 데이터 정제 규칙이다.

표 5. 데이터 정제 규칙(수집 영역)
Table 5. Rules of data cleansing (area of data gathering & cleansing)

유형	정제 규칙
유형1	· 기본값(default)이 정의되어 있는 경우 기본값 세팅 · 기본값(default)이 정의되어 있지 않은 경우 데이터 재 수집
유형2 유형3	· NTIS 표준 스키마 및 코드에 맞게 매핑 되었는지 확인 (매핑 되지 않은 경우 데이터 재 수집)
유형4	· 날짜, 숫자, 단위, ISBN 등 NTIS 표준에 맞는지 확인 · NTIS 표준에 맞지 않는 데이터는 표준 형태로 변형 (연차정보 : 연차만 수집되면, 앞에 '0-'을 붙임) · NTIS 표준으로 매핑되지 않은 경우 데이터 재 수집 · 의미 없는 문자의 반복입력은 삭제(단어, 문장은 살림)
유형5	(5-1) 논문, 특허 저자의 구분자 오류 · 연계기관에서 조치 후 데이터 재수집 · 저자 별 구분자를 표준(;)에 맞게 수정 (5-2) 사업부처명 및 사업부처 코드 오류 <1> 유형1사업부처명을 기초로 해당 코드를 찾아서 코드 기입 <2> 사업부처코드를 기초로 해당 명을 찾아서 사업명 기입 <3> 연계기관에서 조치하도록 피드백

데이터 정제 규칙에 의해 정제가 된다하더라도 일부 오류에 대해서만 가능하다. 전술한 바와 같이 데이터 제공 기관에서 데이터 수정이 없을 경우에는 정제할 수 없는 오류들이 많이 존재한다. 이러한 데이터들은 서비스가 되기에 적합하지 않으므로, 별도의 서비스 기준을 정의하고, 서비스DB 구축 전 서비스 기준에 적합하지 않은 데이터들은 배제한다. 서비스 기준은 서비스를 위해 필수적으로 필요한 최소한의 기준이다. 서비스 건수와 서비스 기준 사이에는 trade-off 관계가 있다. 서비스 기준을 강화하면, 서비스 가능한 건수는 적어진다. 반대로, 서비스 기준은 완화하면, 서비스 가능한 건수는 늘어난다.

서비스 기준은 필수항목과 항목에 존재하는 업무규칙으로 구성된다. 필수항목은 총 14개이며, 사업정보(3개), 과제기본정보(5개), 연구비정보(2개), 역할매핑정보(1개), 참여인력기본정보(2개), 참여인력재직기관(1개)이다. 필수항목 중 7개 항목은 업무규칙을 가지고 있다.(표6)

표 6. 서비스 기준(업무규칙)
Table 6. Rules of data for service

테이블명	항목명	업무 규칙
과제기본정보	당해연도 연구기간	당해연도연구기간은 총연구기간과 같거나 총연구기간 내에 포함되어야 한다.
		당해연도연구기간 시작일의 연도는 기준년도와 동일해야 한다.
	주관연구기관명, 발주기관명	유효한 값이어야 한다. ex) 'ZZ'로 입력된 데이터는 오류
연구비정보	정부투자 연구비	'0' 이상이어야 한다.
	연구비 합계	'0' 보다 커야 한다.
역할매핑정보	역할구분	연구책임자(A)는 과제당 1명 존재해야 한다.
참여인력기본정보	주민등록번호	'XXXXXX-XXXXXXX'이어야 한다.

2. 데이터 분석 및 오류 데이터 예시

수집된 데이터의 오류 분석 결과, 기 정의해 놓은 오류 유형처럼 형식 오류, 코드 및 스키마 오류, 내용 오류 등 다양한 형태의 오류가 검출되었다. 아래는 오류들에 대한 실제 사례이다.

○ 날짜 형식 오류
- 00000000, 19940000, 20059825 등
- 01012004, 05012001, 07012000 등
- 05, 11.24, 10-2005-, 1997-092 등
- 1990712, 2000828, 200505 등
- YYYYMMDDHH24MISS 형식으로 된 경우

○ 숫자 형식 오류
- 원단위 항목의 경우 단위를 "원"이 아닌 백만단위, 천단위로 사용하는 경우
- 100,000, 1,067,000 등
- 210,000,000원, 600,000,000\, 10000000 won, \ 300,000,000 등
- 당해연도 56,119,000원(기업체부담금), 해당없음 등
- .5, 124.12345678 등

○ 텍스트 형식 오류

- -----, ****, ####, 11111, ____
- 없음아ㅣ러만ㄹㅇㄹㅁ
- 입력요망
- 테스트입니다
- 차후 입력요망

○ 정규 형식에 어긋나는 경우

<ISSN_ISBN> 번호

- 10.1002/bmc.792
- 11-1390317-000031-10
- 89-480-0137-X 93520
- DOI 10.1007

3. 오류 데이터 정제

오류 데이터에 대한 정제는 SQL문에 의한 기계적인 배치 정제가 있고, 사람의 판단이 필요한 수작업 정제가 있다. 오류 검출은 대부분 SQL문에 의해 기계적으로 검출이 가능하지만, 정제는 수작업에 의한 정제가 많다. 수작업 정제를 지원하기 위해서는 정제 작업자가 작업할 수 있는 작업 환경을 마련해 주는 것이다. 시스템에서 특정 버튼을 눌렀을 때, 사전에 정의된 오류 유형에 따라 오류가 검출되고, 검출된 오류는 데이터 컬럼별로 오류 표시가 되어 정제 작업자의 입력/수정 UI 화면에 표시된다.(그림3) 정제 작업자는 표시된 오류 데이터에 대해 입력/수정 UI 화면을 통해 데이터를 정제할 수 있다. 정제 시 기준데이터가 있는 경우에는 기준 데이터를 기준으로 오류를 정정할 수 있고, 기준데이터가 없는 경우에는 사전에 정의된 정제 지침에 의해 정제를 수행하였다.

그림 3. 데이터 정제시스템(정제 작업자 화면)
Figure 3. Data cleansing System(UI for workers)

4. 데이터 정제 결과

NTIS 데이터의 정제는 2단계로 이루어진다. 초기 데이터를 보유하고 있는 연계 기관들에서 NTIS 표준에 따라 1차로 정제를 한 후 수집된 데이터에 대해서 2차 정제를 수행한다. 2차 정제는 각 기관에서 수집된 데이터에 오류가 있는지 확인하고, 오류 사항에 대해서는 각 기관에 피드백하거나 내용수정이 아닌 경우에 한해서 직접 정제(2차 정제)를 수행한다. 분석 후 각 기관에서 수정하도록 하는 것이 기본 원칙이다.

기계적 정제 및 수작업 정제의 효과를 측정하는 방법에는 2가지가 있다. 첫째는 최초 수집된 데이터의 품질과 정제 후 서비스되는 데이터의 품질을 측정하여 두 시점의 품질을 비교하는 방법이고, 둘째는 수집 시점의 데이터만 분석하되, 각 기관에 피드백 하기 전 데이터의 품질과 피드백을 반복하여 각 기관에서 데이터를 정제한 후의 품질을 비교하는 방법이다.(그림4)

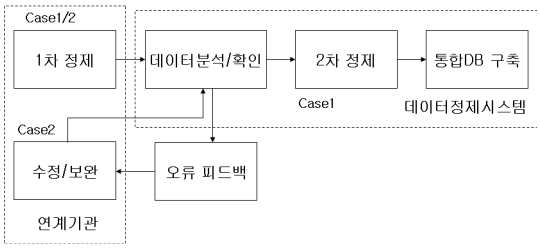

그림 4. 데이터 정제 단계 및 결과 측정 방안
Figure 4. Data cleansing steps and measuring the result in NTIS

본 연구에서는 Case2 방식을 통해 데이터 정제 결과를 측정하였다. NTIS사업에서는 오류 부분에 대해서 직접 정제하지 않고, 데이터 소스를 제공하는 연계 기관에 데이터 정제를 의뢰함으로써 정제의 정확도를 높이고자 하였다. 따라서 Case1 방식이 일반적일 수 있지만, 데이터 오류에 대한 피드백 과정이 있으므로 2차 정제를 통해 데이터 품질이 높아지는 경우는 적다고 볼 수 있기 때문에 Case2 방식도 유효하다고 판단할 수 있다.

Case2 방식은 초기 데이터를 제공하고, 제공된 데이터에 대한 분석에 따른 피드백이 이루어진 기관들에 적용이 가능하다. 8개 기관에 대해서 Case2 방식이 적용 가능하였다.

데이터 품질 측정 방식은 한국데이터베이스진흥원에서 제시한 '데이터 품질진단 절차 및 기법(Ver1.0)'에 의해 실시하였다[16]. 데이터 품질진단 절차는 품질측정계획 수립, 품질

측정 체크리스트 준비, 데이터 품질측정 수행, 데이터 품질측정 결과보고 등 4단계로 진행된다. 본 연구에서는 데이터 품질진단 절차대로 수행하였고, 품질측정 결과만 언급하고자 한다.

품질측정의 기준은 NTIS에서 정의해 놓은 업무규칙이고, 업무규칙에 위배된 데이터는 오류로 규정하였다. 여기서 의미하는 데이터는 레코드(Record) 단위가 아니라 필드(Field) 단위이다. 하나의 레코드에는 여러 개의 개별 필드가 존재한다. 오류 측정식은 아래와 같다.

☞ 오류율(Error Rate) = 오류컬럼수* / 점검대상컬럼수**

* 테이블1[오류컬럼수]+테이블2[오류컬럼수] + ... + 테이블N[오류컬럼수]

** 테이블1[레코드수×컬럼수]+테이블2[레코드수×컬럼수] + ... + 테이블N[레코드수×컬럼수]

위 오류 측정식에 의해 8개 기관을 대상으로 오류율을 측정하였다.(표7) 전체적으로 정제 후 데이터(오류율 4.58%)가 정제 전 데이터(11.23%)보다 품질이 좋아졌음을 알 수 있다. 기관별로도 2개 기관(E, G)을 제외하고는 대부분 정제 후 데이터가 품질이 높은 것으로 나타났다. 품질이 더 나빠진 2개의 기관은 피드백 결과에 대해서 조치가 제대로 이루어지지 않았기 때문에 결과치로 정제의 효과를 측정하기에는 어려움이 있다. 2개 기관을 제외한 결과는, 11.48%(정제 전)에서 4.52(정제 후)로 역시 데이터 정제 후 데이터 품질이 크게 향상되었음을 알 수 있다.

B기관의 경우는 피드백이 제일 많이 일어났고, 실제 오류율이 반으로 줄었다. 절대적으로 보면 타 기관들보다 오류율이 많이 높지만, 중복 및 불필요한 데이터들이 정제되었고, 정제의 효과가 오류율의 감소로 나타났음을 알 수 있다.

표 7. 데이터 품질 측정 결과
Table 7. The result of measuring data quality

기관	초기데이터			정제 후 데이터			피드백 횟수**
	측정 컬럼수*	오류 컬럼수	오류율 (%)	측정 컬럼수	오류 컬럼수	오류율 (%)	
A	50,682	696	1.37	35,122	92	0.26	2회
B	420,325	76,943	18.30	151,639	11,138	7.35	5회
C	83,215	1,782	2.14	34,904	512	1.47	2회
D	60,457	634	1.05	31,703	6	0.02	1회
E	46,211	935	2.02	28,770	896	3.11	2회
F	98,926	6,958	7.03	60,718	2,719	4.48	2회
G	8,059	544	6.75	40,573	2,765	6.81	1회
H	181,874	18,181	10.00	32,430	932	2.87	2회
계	949,749	106,673	11.23	415,859	19,060	4.58	-

* 점검대상컬럼수(레코드수 × 컬럼수)
** 오류 데이터에 대한 수정 횟수

V. 결론

본 연구의 목적은 분산되어 있는 국가R&D 데이터베이스들을 공동 활용하기 위해 데이터베이스 통합 모형을 개발하고, 통합 모형에 기초해 통합된 데이터베이스 내의 데이터의 품질을 제고하기 위한 방안을 제안하는 것이다. 이를 위해 2장에서는 데이터 정제에 관한 이론적 고찰을 하였고, 3장에서는 데이터베이스 통합 및 관련 개념들을 정리하였다. 4장에서는 NTIS 적용 사례를 통해서 구체적인 데이터베이스 통합 모델을 제시하였고, 데이터 품질 제고를 위해 통합된 데이터베이스 내의 데이터 정제의 필요성을 제기하였다. 실제 데이터 오류율 측정을 통해 데이터 정제가 데이터 품질 제고에 기여함을 입증하였다.

본 연구는 다음과 같은 시사점을 준다.

첫째, 분산되어 있는 데이터베이스를 통합하여 통합 데이터베이스 구축을 위해서는 데이터 통합 모델 설계가 중요하다. 특히, 국가R&D정보 데이터의 실시간 서비스를 위해 데이터 품질을 고려한 데이터베이스 통합 개념이 필요하다. 이를 위해 데이터 정제가 동시에 이루어질 수 있는 데이터베이스 통합 절차 및 모델이 필요하다.

둘째, 데이터 통합에서 필수적으로 제기되는 데이터 품질 문제 해결을 위해 데이터 정제가 반드시 필요하며, 데이터 정제를 위해서는 업무규칙 정의를 통한 정제 기준(표준 데이터)을 마련하고, 데이터분석을 통한 오류 데이터 검출, 그리고 데이터 수정 과정이 필요하다.

셋째, 데이터 품질측정 기준에 의한 데이터 오류율 측정을 통해 본 연구에서 제시한 데이터 통합 모델에 기초하여 데이터 정제가 수행된 데이터는 그렇지 않은 데이터보다 품질이 향상되었음을 알 수 있다.

본 연구는 NTIS의 사례만을 조사한 것이어서, 다른 사례에 대한 비교가 어렵다는 한계점을 가지고 있다. 현재 국내에서 국가적으로 범부처의 R&D정보를 수집하여 통합 데이터베이스를 구축하는 사례를 찾아보기 어려웠다. 민간 기업에서는 기업 간 합병이 이루어져 기업정보 데이터베이스가 통합되는 경우는 있다. 하지만, 국가R&D정보 특성 상 민간 기업의 사례를 그대로 적용하는 것은 문제점이 있어 보인다. 향후에는 해외 사례를 더 많이 분석하여 NTIS에 적용할 수 있는 연구가 필요할 것으로 판단된다.

참고문헌

[1] A. Levitin and T. Redman "A Model of the Data (life) cycles with application to quality," Information and Software Technology, Vol. 35, No. 4, pp. 217-223, Apr. 1993.

[2] E. Simoudis et al, "Using Recon for Data Cleaning," KDD-95 Proceedings, pp. 282-287, 1995.

[3] I. Guyon et al, "Discovering Informative Patterns and Data Cleaning," AAAI-94 Workshop on Knowledge Discovery in Databased, AAAI Technical Report WS-94-03, pp. 145-156, Mar. 1996.

[4] R. Kimbal, "Dealing with Dirty Data: Every serious data warehouse application needs good data, yet few people address the issue", DBMS, Vol. 9, No. 10, pp. 55-62, Sep. 1996.

[5] M. A. Hernandez and J. S. Stolfo, "Real-World Data is Dirty: Data Cleansing and The Merge/Purge Problem," Journal of Data Mining and Knowledge Discovery, Vol. 2, No. 1, pp. 9-37, Jan. 1998.

[6] H. Galhardas et al, "An Extensible Framework for Data Cleansing," Rapport Recherche, Institute National de Recherche en informatique et en Automatique, Jul. 1999.

[7] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Bulletin of the Technical Committee on Data Engineering, Vol. 23, No. 4, pp. 3-13, Dec. 2000.

[8] A. D. Chapman, "Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data," Global Biodiversity Information Facility, Jul. 2005.

[9] J. I. Maletic and A. Marcus, "Data Cleansing: Beyond Integrity Analysis," Proceedings of the Conference on Information Quality, pp. 200-209, Jun. 2000.

[10] H. J. Whang, "A Study on Data Cleansing Methodology," Baewha Women's Univ., Vol. 23, pp. 185-203, May 2004.

[11] K. R. Shon, "A Data Quality Improvement Method in Integrations of Distributed Data : National Science & Technology Information Services," The Journal of Korean Institute of Marine Information and Communication Sciences, Vol. 13, No. 8, pp. 1623-1636, Aug. 2009.

[12] J. A. Seol, "Design of Data Integrating System Using XML Metadata Registry in a Distributed Environment", Kwangwoon Univ., Feb. 2004.

[13] Jae-Soo Kim, "Introduction of NTIS," Journal of Scientific & Technological Knowledge Infrastructure, Vol. 30, pp. 31-34, May 2008.

[14] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," ACM Computing Surveys (CSUR) - Special issue on heterogeneous databases, Vol. 22, No. 3, pp. 183-236, Sep. 1990.

[15] I. N. Kwon et al., "R&D Information Distribution Infrastructure," Journal of scientific & technological knowledge infrastructure, Vol. 30, pp. 45-53, May 2008.

[16] Data Quality Assessment Procedure Manual(Ver1.0), Korea Database Agency, Oct. 2009

저 자 소 개

신 성 호

2000 : 경북대학교 경영학과 경영학사.

2002 : 경북대학교 경영학과(MIS전공) 경영학석사.

현 재 : 한국과학기술정보연구원 선임연구원

관심분야 : 데이터통합, 데이터품질, IS 평가

Email : maximus74@kisti.re.kr

윤 영 준

1997 : 충남대학교 문헌정보학과 도서관학학사.

2001 : 충남대학교 문헌정보학과 도서관학석사.

현 재 : 한국과학기술정보연구원 선임연구원

관심분야 : DB구축, 데이터정제, 인력정보

Email : yjyoon@kisti.re.kr

양 명 석

1999 : 충남대학교 컴퓨터과학과 이학학사.

2001 : 충남대학교 컴퓨터과학과 이학석사.

현 재 : 한국과학기술정보연구원 선임연구원

관심분야 : 정보검색, 데이터마이닝, DB

Email : msyang@kisti.re.kr

김 진 만

2002 : 동의대학교 컴퓨터공학과 공학사.

2004 : 동의대학교 컴퓨터공학과 공학석사.

2007 : 동의대학교 컴퓨터응용공학과 공학박사

현 재 : 한국과학기술정보연구원 선임연구원

관심분야 : 데이터품질, 네트워크

Email : iricejm@kisti.re.kr

손 강 렬

1999 : 국립공주대학교 전자계산학과 공학석사.

2009 : 국립공주대학교 컴퓨터공학과 공학박사.

현 재 : 한국과학기술정보연구원 책임연구원

관심분야 : 정보관리, 데이터품질, 인력정보

Email : krshon@kisti.re.kr