분산 데이터의 통합시 데이터의 품질향상 방안: 국가과학기술종합정보시스템

손강렬*

A Data Quality Improvement Method in Integrations of Distributed Data: National Science & Technology Information Services

Kang-Ryul Shon*

요 약

현재 국내의 정부 R&D 사업은 300여개에 이르고 있고, 이를 사업의 특성별로 16개 국가R&D 관련 부처 · 청의 15개 대표연구관리전문기관에서 각각 관리하고 있다. 이로 인하여 발생하는 국가 R&D에 대한 중복 투자와 체계적인 R&D연구과제 및 성과관리의 미흡으로 R&D 투자의 효율성에 대한 문제가 계속해서 제기되고 있다. 그러한 가운데 이러한 문제를 해결하기 위하여 교육과학기술부는 국가연구개발의 기획에서 성과활용에 이르기까지 연구개발의 효율화를 지원할 수 있는 국가 R&D 포털시스템으로써 국가과학기술종합정보시스템(NTIS)을 구축하고 있다. NTIS와 같이 분산된 데이터의 통합시 동일한 의미의 데이터들이 각 조직에서 달리 명명되고 다른 데이터 유형으로 되어 있기에 통합된 데이터의 정확성과 높은 수준의 품질을 달성하는 것이 어려운 문제이다. 본 논문에서는 NTIS 시스템의 인력/과제/성과 정보의 통합DB 구축 및 연계방식과 이를 통해 수집된 데이터의 품질관리를 위한 데이터 정제 프로세스를 고찰해 본다. 그 과정에서 발생할 수 있는 데이터 품질문제의 요인을 분석하여 NTIS의 데이터 품질향상을 위한 개선방안을 제시한다.

ABSTRACT

A currently domestic governmental R&D business is early to 100. And this is each managed individually in 15 professional organizations of research and management by characteristics of a business. For this Reason, A redundant investment issue regarding national R&D occurs, and an issue regarding efficiency of R&D investment by insufficiency of systematic R&D research project and result management is continuously raised. Ministry of Education Science and Technology establishing National Science & Technology Information Service(NTIS) in order to solve these issues. NTIS is the national R&D Portal System which can support efficiency of research and development to result utilization in planning of national research and development. As data of the same meaning are named particularly in each organizations, and that made to different data types, It is an issue to be difficult to achieve high level quality, accuracy of integrated data in case of integration of distributed data like NTIS In this paper We consider integrated DB constructions and Information Linking of R&D Participants/Projects/Results information in a NTIS system for data quality Improvement. and then We analyze the cause of the data quality problem, and we propose the improvement plan for data quality elevation of NTIS system.

키워드

데이터베이스통합, 데이터품질관리, 데이터정제, 데이터품질문제분석

* 한국과학기술정보연구원(KISTI) 책임연구원

접수일자 2008. 12. 29
심사완료일자 2009. 01. 16

Ⅰ. 서 론

우리나라 정부 R&D 예산이 최근 몇 년간 두 자릿수를 넘는 높은 증가율을 기록하고 있다. 그러한 가운데 국가 R&D 사업이 점차 확대되면서 R&D 투자 효율성에 대한 관심이 어느 때보다 높아지고 있다. 최근에는 정부 R&D 예산이 사상 최초로 10조원을 넘는 등 그 규모는 점점 더 커지고 다양화되고 있다. 그러나 우리나라 R&D 투자의 효율성은OECD 평균수준이나, 미국, 일본보다 낮아 R&D 투자 효율성 제고에 대한 노력이 필요한 상황이다. 특히 국가R&D 사업에 대한 정보가 정부 부처별, 기관별로 관리되면서 발생하는 국가 R&D에 대한 중복투자 등의 문제 해결과 함께 R&D사업에서 산출된 정보와 자원을 적극적으로 활용하여 연구개발 효율성과 생산성을 제고하자는 요구사항이 증가하였다. 이에 교육과학기술부는 R&D 관련 부처 · 청의 대표연구관리전문기관(이하 과제관리기관)과 연계하여 정보를 수집 · 가공한 후, 공동 활용함으로써 R&D 투자 효율성을 제고하기 위한 과학기술종합정보서비스(이하 NTIS)를 구축하였다. <그림 1>은 국가R&D정보 지식포털(NTIS) 개념도로서 국가연구개발 사업 및 과제가 어떻게 수집되어 활용 및 서비스 되는지를 보여주고 있다[1].

국가R&D 관련 정보를 공동활용하기 위해서는 국가R&D정보의 통합 데이터베이스 구축이 필요하다. NTIS는 국가 R&D 정보의 통합DB 구축을 위하여 각 기관 특성별로 기 구축된 R&D 정보관리 시스템은 그대로 유지하고 이들 기관과 공통으로 활용할 수 있는 데이터 표준스키마를 정의하여 NTIS 통합DB를 구축하고 있다. 그리고 과제관리기관의 레거시 DB와 통합DB간의 정보연계를 통하여 국가R&D 관련 정보를 공유하고, 공동 활용할 수 있도록 하고 있다. 이러한 NTIS가 그 활용도를 높이고, 국가 연구개발의 효율성과 생산성을 극대화시키기 위해서는 무엇보다 NTIS가 제공하는 데이터의 품질이 가장 중요한 요소가 될 것이다.

따라서 본 논문에서는 선행연구로서 다양한 데이터베이스 통합 방법들과 통합된 데이터의 품질향상 및 품질관리 방법들을 고찰해보고 NTIS 통합DB 구축에 따른 품질제고 방안을 제시한다. 이를 위하여 NTIS의 국가R&D참여인력/과제/성과정보에 대한 통합DB 구축을 위한 정보연계 방식과 데이터 정제 프로세스를 살펴보고 여기서 통합된 데이터베이스 환경에 따른 데이터 품질관리관점에서 데이터의 품질에 영향을 미칠 수 있

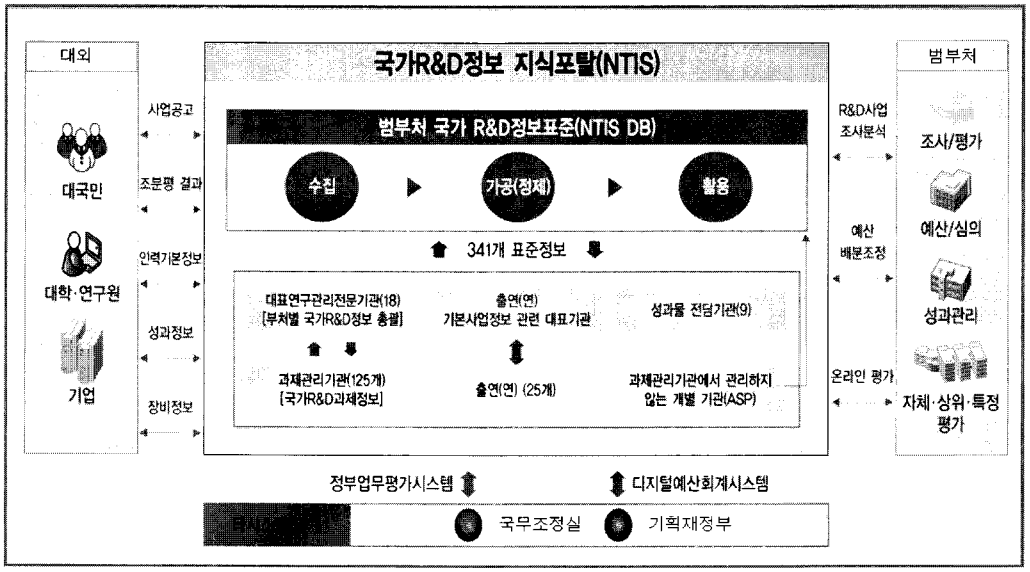

그림 1. 국가R&D정보 지식포털(NTIS) 개념도
Fig. 1 Concept of National R&D Information Knowledge Portal(NTIS)

는 요인에 대한 분석을 통하여 그에 대한 개선방안을 제안한다.

II. 선행연구

2.1 다양한 데이터베이스 통합 방법

지금까지 이질적인 데이터베이스간 통합을 위한 많은 방법들이 제안되었다. 이들은 일반적으로 데이터 모델과 스키마의 이질성을 극복하기 위해 GDM (Global Data Model)과 전역스키마를 제공해야 하며, 사용자의 정보서비스요청(질의)에 대한 부질의(Subquery)집합으로 변환, 그리고 전역 스키마와 지역 스키마간 번역 등 복잡한 구조를 기반으로 하고 있다.

이는 최종 사용자에게 다양한 질의(정보)와 이질적인 지역 데이터베이스들에 대한 단일의 접근방법을 제공하기 위함이다[2]. 여기서 정보 시스템 통합방법의 3가지 유형으로 살펴보도록 한다. 첫째, 고전적 접근 방식이 있다. 이 방식은 다양한 지역 소스(local source) 스키마의 차이를 해결하기 위한 하나의 전역 스키마 작성이 핵심이 된다. 즉, 이질적인 지역소스의 스키마를 바탕으로 전역 스키마를 작성하고, 이를 통해 시스템 사용자는 다양한 지역 소스에 대한 일정한 형태(전역 스키마)로의 접근이 가능하며, 다양한 지역 소스가 통합되는 효과를 준다. 하지만 지역소스가 새로이 추가되거나 수정되면 지역 소스를 바탕으로 전역 스키마의 갱신이 요구되는 단점이 있다[3,4]. 둘째, 고전적 접근방식이 유발했던 데이터베이스 사용자 이질성의 문제를 해결한 연합된 접근 방법이 있다. 하지만, 여전히 전역 스키마의 방식을 사용하고 있고, 전역 스키마 또한 여전히 정적이기 때문에 지역 소스의 추가나 수정시 전역 스키마의 갱신이 요구되어진다[5]. 셋째, 분산 객체관리방식이 있다. 이는 분산되고 이질적인 데이터베이스를 분산 객체 공간의 객체 집합 모델 기반으로서, 연합된 접근 방법을 일반화 시켰다. 이것은 공통 객체 모델과 공통 객체 질의 언어를 바탕으로 하고 있다[6].

이러한 분산된 데이터의 통합과 관련한 또 다른 연구로서 이윤준[7]의 연구에서는 분산 데이터의 통합된 환경에서 품질향상을 위한 워크플로우상에서의 효과적인 작업할당 방법을 제안하고 있다.

2.2 데이터 품질의 개념과 특성

Larry P. English는 데이터품질(Data Quality)의 정의를 기업과 고객의 목표를 달성하기 위해 데이터에 대한 이해관계자의 기대를 충족시키는 것이라고 정의하였다. 또한, 데이터 품질유지를 위한 원칙으로 과학적 기법을 통해 고객에 집중하여 데이터에 대한 개선활동을 수행하도록 제시하고 있다.[8]

데이터 품질특성(Data Quality Characteristics)은 소프트웨어 품질 특성[9]과는 달리 표준이 명확히 정립되어 있지 않고, 각기 필요성에 따라 조금씩 연구가 진행되어 왔다. 그 중 대표적인 것으로 Wang의 연구[10]를 들 수 있는데, 데이터품질은 4가지 차원, 즉 정확성(accuracy), 적시성(timeless), 완전성(completeness), 일관성(Consistency)으로 구분된다.

각 데이터 품질특성에 대한 측정은 소프트웨어 품질 측정 표준인 ISO/IEC 9126을 기반으로 데이터 품질을 측정하기 위한 방안에 대한 연구가 진행되고 있다. 최병주[11]는 오류데이터를 분류하고, 그것으로부터 데이터 품질 특성을 측정하기 위한 메트릭을 제시하였다.

2.3 데이터 품질관리

데이터품질관리[12]란 기관이나 조직 내외부의 정보시스템 및 DB 사용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 및 개선활동을 의미한다. 데이터품질관리에 대한 초기 연구는 품질 측정에 대한 현상분석 중심이었으나, 점차적으로 품질개선을 위한 모델 중심으로 바뀌고 있다.

MIT의 TDQM(Total Data Quality Management) 프로그램[10]은 데이터품질을 본질적(Intrinsic)품질, 연관적(Contextual) 품질, 표현적(Representational) 품질 등의 카테고리로 분류하고, 각 카테고리 별로 데이터품질 문제가 발생하는 패턴 및 개선방안을 연구하고 있다.

Larry P. English의 TIQM(Total Information Quality Management) 모델[8]은 6단계의 프로세스로 구성되어 있으며, 각 단계의 프로세스들은 평가, 유지보수, 데이터 이행 통제, 유지보수를 위한 프로세스 개선 및 조직을 데이터 품질의 문화로 전환하는 프로세스로 구성된다. 한국데이터베이스진흥센터는 데이터 품질관리 지침[12]을 통해 데이터의 품질을 개선할 수 있는 프레임워크를 제시하였다. 데이터품질관리 프레임워크는

Enterprise Architecture의 개념을 도입한 것으로 표준데이터를 정보시스템의 데이터 품질 확보를 위한 필수 요소로 정의하고 있다. 한국데이터베이스진흥센터가 제시한 품질평가 모형에서는 데이터베이스 품질을 데이터 품질, 시스템 품질뿐만 아니라, 데이터 관리 프로세스 품질을 포함하도록 확장하였다. 즉, 데이터 값, 데이터 구조의 품질을 포함하는 데이터 자체의 품질과 이들 데이터를 다루는 정보시스템의 품질뿐만 아니라, 이들 데이터를 관리하는 프로세스의 품질을 포괄적으로 다루고 있다[13].

데이터의 품질관리 프로세스의 중요한 과정중의 하나는 데이터 정제이다. 데이터 정제에 대한 명확한 정의가 있지 않지만 일반적으로 오류가 있는 데이터를 검출하거나, 값이 입력되지 않은 데이터의 검출, 중복된 데이터 검출 등 여러 형태의 데이터 오류와 불일치를 검출하고 제거하는 과정을 데이터 정제라고 한다. 이러한 데이터 정제는 다양한 데이터를 통합 또는 대용량의 데이터 품질을 관리하는데 있어서 그 중요성은 날로 커지고 있다.

일관성 없고 불완전하며 오류가 있는 데이터는 데이터의 무결성과 질을 보장하기 위하여 데이터 정제 과정을 거쳐야 하며, 데이트의 스키마와 데이터표준 등 여러 요소가 복합적으로 고려되어 진행되어야 한다[14].

Ⅲ. NTIS의 국가R&D정보 통합DB 구축을 위한 정보연계 방식

본 장에서는 NTIS의 국가R&D참여인력/과제/성과정보의 통합DB 구축을 위한 정보연계 방식을 살펴보고 품질관리 측면에서의 문제점을 분석하고자 한다.

3.1 NTIS 국가R&D참여인력정보의 정보연계

NTIS의 국가R&D참여인력정보는 국내의 여러 과제관리전문기관에 분산되어 구축・운영 중인 과학기술인력DB와 연계하여 국가R&D사업에 참여하는 연구책임자 및 연구자에 대하여 개인동의를 받아 통합DB를 구축하고, 통합 검색 서비스 및 각종 현황정보, 사업/과제/성과정보와 연계한 사용자 중심의 국가 R&D종합정보 서비스를 제공하고 있다.

NTIS의 국가R&D참여인력정보의 통합DB 구축 및 데이터 연계방법은 먼저 각 과제관리전문기관이 기존에 보유하고 있는 과학기술인력DB를 모두 수집하는 단계를 거치고, 수집된 데이터에 대한 개인동의 절차를 거쳐서 개인동의를 획득한 데이터를 표준스키마 및 코드매핑 작업을 거쳐서 NTIS 통합인력DB를 구축하게 되며, 데이터 정제이후에 서비스를 제공하게 된다.

<그림 2>는 NTIS의 국가R&D참여인력정보의 정보연계 방식을 나타내고 있다. 다음은 <그림 2>를 토대로 정보연계에 따른 주요 핵심요소를 설명한다.

*연계기관*은 과제관리기관으로서 기관에서 기 구축된 *기관DB(레거시)*와 NTIS와 연계하기 위한 *정보연계서버*가 있다. 정보연계서버는 NTIS공동활용연계DB(인력/과제/성과), 인력정보DB, SIMS DB로 구성된다.

NTIS공동활용연계DB는 연계기관의 R&D인력/과제/성과 데이터에 대하여 NTIS와 연계할 수 있는 DB스키마(수집/연계)를 가지고 있으며, 기관DB에서 신규/변경된 정보 발생 시 정보연계 정책에 따라 국가R&D정보의 수집/연계 창구역할을 한다. ***인력정보DB***는 NTIS와 정보연계를 통하여 NTIS 통합홈페이지 또는 타 과제관리기관에서 변경/수정된 내용이 반영될 수 있도록 하여 인력정보에 대한 최신성을 유지하도록 한다. ***SIMS DB***는 과제관리기관에서 과제/성과 데이터를 입력/수정/보완할 수 있는 창구역할을 하며, 또한 SIMS를 통해 과제/성과 데이터 입력시 국가R&D인력정보(세부항목포함)를 입력/수정/보완할 수 있는 창구역할을 한다.

그리고, ***NTIS의 데이터 흐름***은 크게 3개의 부분으로 나누어진다. 첫째는 NTIS연계 SEED DB부분, 둘째는 NTIS관리DB부분, 셋째는 NTIS서비스DB 부분이다.

첫째 ***NTIS연계 SEED DB***는 연계기관의 정보연계서버로부터 수집된 국가R&D참여인력/과제/성과정보를 NTIS로 전송받는 역할을 하고, 국가R&D참여인력의 경우 SEED DB의 View를 통하여 연계기관의 정보연계서버에 있는 인력정보DB로 연계를 통하여 수정/변경된 국가R&D참여인력 데이터를 전송하는 역할을 한다. 둘째 NTIS관리DB는 NTIS 표준스키마에 따라 구축된 ***NTIS통합DB***영역으로써 국가R&D참여인력정보를 통합 관리하는 *통합인력관리DB*와 국가R&D과제/성과정보를 통합 관리하는 *과제/성과 통합SEED*와 *과제/성과 관리DB*로 구성된다. 셋째 ***NTIS서비스DB***는 NTIS관리DB에 통합 구축된 NTIS 국가R&D참여인력/과제/성과 데이터

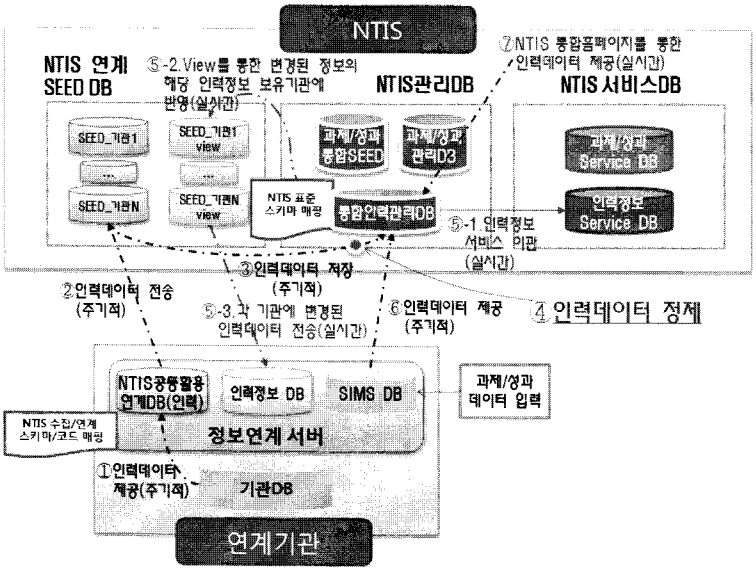

그림 2. 연계기관과 NTIS 국가R&D참여인력정보 통합DB와 데이터 연계구조
Fig. 2 Information Linking Structure of R&D Project Participants Information between NTIS Integration DB and Associated Institutions DB

를 최종 서비스하는 역할을 한다. ***NTIS서비스DB**는 인력정보 **Service DB**와 과제/성과 **Service DB**로 나뉘어져 있다.*

국가R&D참여인력정보에 대한 정보연계 프로세스는 다음과 같다.

① 인력데이터 제공(주기적) → 연계기관 기관DB에서 정보연계서버의 NTIS공동활용연계DB로 데이터 제공, NTIS 수집/연계 + 스키마 및 코드매핑 수행
② 인력데이터 전송(주기적) → 연계기관의 정보연계서버에서 NTIS 연계 SEED DB로 데이터 전송
③ 인력데이터 이관(주기적) → NTIS 연계 SEED DB에서 통합인력관리DB로 데이터 전송, NTIS 표준 스키마/코드 매핑 수행
④ 인력데이터 정제(실시간) → ③에서 저장된 데이터에 대하여 정제 프로세스 수행
⑤ ※데이터 정제가 완료됨과 동시에 다음의 ⑤-1, ⑤-2,⑤-3이 일률적으로 진행됨
⑤-1. 인력정보 서비스 이관(실시간) → 정제가 완료된 인력 데이터를 인력정보 Service DB로 이관 수행
⑤-2. View를 통한 변경된 정보의 해당 인력정보를 보유한 기관에 반영(실시간)
⑤-3. 각 기관에 변경된 인력 데이터 전송(실시간)
⑥ 인력데이터 제공(주기적) -> 연계기관의 SIMS를 통해 국가R&D과제/성과 데이터 입력시 국가R&D인력정보(세부항목포함) 입력 창구를 통하여 NTIS 통합인력관리DB로 인력데이터를 직접입력 받음
⑦NTIS 통합홈페이지를 통한 인력데이터 제공(실시간)

3.2 NTIS 국가R&D과제/성과정보의 정보연계

NTIS의 국가R&D과제/성과정보의 통합DB 구축 및 데이터 연계방법은 먼저 각 과제관리전문기관이 기존에 보유하고 있는 국가R&D과제/성과정보를 모두 수집하는 단계를 거치고, 수집된 데이터를 수집/표준 스키마 매핑 작업을 거쳐서 NTIS 통합 국가R&D과제/성과DB를 구축하게 된다. 구축된 국가R&D과제/성과정보는 데이터 정제 이후에 서비스를 제공하게 된다.

<그림 3>은 국가R&D과제/성과정보의 정보연계방식을 나타내고 있다. 다음은 <그림 3>을 토대로 국가R&D과제/성과정보의 정보연계에 따른 주요 핵심요소를 설명한다.

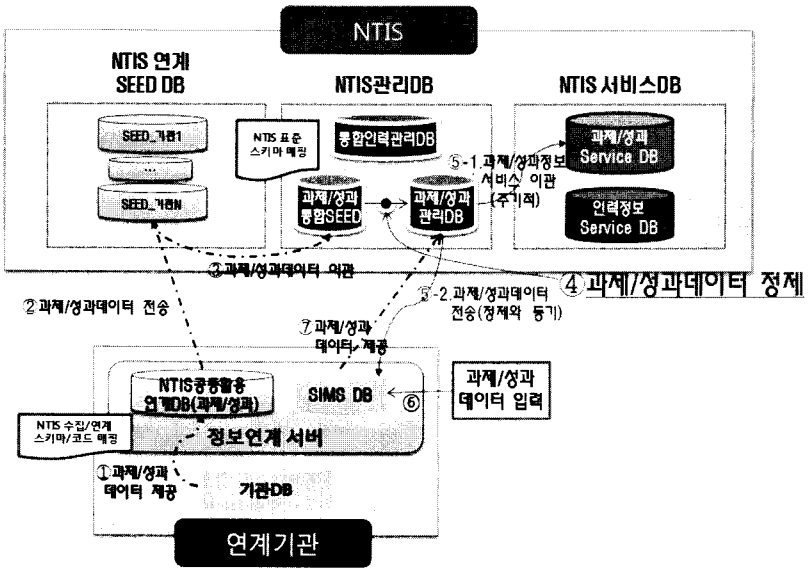

그림 3. 연계기관과 NTIS 과제/성과정보 통합DB와 데이터 연계구조
Fig. 3 Information Linking Structure of R&D Projects/Results Information between NTIS Integration DB and Associated Institutions DB

*연계기관*은 과제관리전문기관으로서 기관에서 기구축된 *기관DB(레거시)*와 NTIS 연계를 위한 정보연계서버가 있다. 정보연계서버는 ***NTIS공동활용연계DB*** (인력/과제/성과), SIMS DB로 구성된다. NTIS공동활용연계DB는 연계기관의 R&D인력/과제/성과정보에 대하여 NTIS와 연계할 수 있는 DB스키마(수집/연계)를 가지고 있으며, 정보연계 정책에 따라 주기적으로 국가R&D 정보를 수집/연계하는 창구역할을 한다. 그리고 ***SIMS DB***는 과제관리기관에서 국가R&D과제/성과정보를 입력/수정/보완할 수 있는 창구 역할을 한다.

그리고 ***NTIS의 데이터 흐름***은 크게 3개의 부분으로 나누어진다. 첫째는 NTIS연계 SEED DB부분, 둘째는 NTIS관리DB부분, 셋째는 NTIS서비스DB 부분이다.

첫째 ***NTIS연계 SEED DB***는 연계기관의 정보연계서버로부터 수집된 국가R&D참여인력/과제/성과정보를 NTIS로 전송받는 역할을 한다. 둘째 ***NTIS관리DB***는 NTIS 표준스키마에 따라 구축된 NTIS 통합DB영역으로써 국가R&D참여인력정보를 통합 관리하는 *통합인력관리DB*와 국가R&D과제/성과정보를 통합 관리하는 *과제/성과 통합SEED*와 *과제/성과 관리DB*로 구성된다. *과제/성과 통합SEED*는 NTIS연계 SEED DB의 국가R&D과제/성과 데이터를 NTIS 표준스키마에 따라 통합시킨 DB로써 이를 대상으로 데이터 정제과정을 수행한다. 데이터 정제과정이 완료된 데이터는 *과제/성과 관리DB*로 저장되고 정책에 따라 ***NTIS 과제/성과 Service DB***로 최종 이관되고, ***SIMS DB***와 실시간으로 동기화 된다. 셋째 ***NTIS 서비스DB***는 NTIS관리DB에 통합 구축된 NTIS 국가R&D참여인력/과제/성과 데이터를 최종 서비스하는 역할을 한다. ***NTIS서비스DB***는 *인력정보 Service DB*와 과제/성과 **Service DB**로 나뉘어져 있다.

국가R&D과제/성과정보는 다음의 특징을 가지고 있다. 각 과제관리기관의 정책에 따라 발생한 국가R&D과제/성과 데이터를 입력하게 되고 인력정보처럼 수시로 변경/수정이 이루어지지 않는 특성을 가지고 있다. 그리고 최종적으로 확정된 과제/성과 정보는 다시 변경되거나 삭제되면 안 된다는 특성을 가진다.

국가R&D과제/성과정보에 대한 정보연계 프로세스는 다음과 같다.

① 과제/성과데이터 제공(주기적) → 연계기관 기관DB에서 정보연계서버의 NTIS공동활용연계DB로 데이

터 제공, NTIS 수집/연계 및 스키마/코드 매핑 수행

② 과제/성과데이터 전송(주기적) → 연계기관의 정보연계서버에서 NTIS 연계 SEED DB로 데이터 전송

③ 과제/성과데이터 이관(주기적) → NTIS 연계 SEED DB에서 과제/성과통합SEED로 데이터 이관, NTIS 표준 스키마/코드 매핑 수행

④ 과제/성과데이터 정제(주기적) → 과제/성과통합 SEED에 저장된 데이터에 대하여 정제 프로세스를 수행하고, 정제가 완료된 데이터는 과제/성과관리 DB로 최종 저장됨

⑤ 과제/성과관리DB에 저장된 데이터의 이관(주기적) → 데이터 이관에 따라 다음의 ⑤-1,⑤-2가 일률적으로 진행됨

⑤-1. 과제/성과정보 서비스 이관 → 정제가 완료된 인력 데이터를 인력정보 Service DB로 이관 수행

⑤-2. 과제/성과 데이터 전송 → 과제/성과관리DB에서 연계기관의 SIMS DB와 동기화

⑥ 과제/성과 데이터 입력

⑦ 과제/성과데이터 제공(주기적) → SIMS DB를 통하여 ⑥에서 입력된 데이터를 과제/성과관리DB로 직접 제공

Ⅳ. NTIS의 데이터 품질향상을 위한 데이터 정제 프로세스

본 장에서는 NTIS의 국가R&D참여인력/과제/성과 데이터에 대한 품질향상 절차 중의 하나인 NTIS의 데이터 정제 프로세스를 살펴보도록 한다.

4.1 NTIS의 국가R&D참여인력 데이터 정제

국가R&D참여인력 정보는 과학기술 관계 장관회의에서 확정된 국가R&D참여인력정보 관련 57개 항목(속성)을 관리하고 있다. 그리고 15개 과제관리기관의 데이터를 Bulk(Off-Line) 또는 정보연계(On-Line) 형태로 수집하여, NTIS 인력데이터 정제 지침 및 매뉴얼에 따라 프로그램에 의한 기계작업과 정제 작업자에 의한 수작업으로 데이터 정제를 수행하고 있다. <그림 4>는 NTIS의 국가R&D참여인력에 대한 데이터 정제업무 프로세스를 나타내고 있다.

데이터의 정제작업은 NTIS인력정보 데이터 정제 지침에 따라 수집된 국가R&D참여인력정보 데이터를 NTIS의 표준코드와 매핑하고 가비지(garbage) 데이터에 대한 처리를 수행한다. 그리고 이중 등록된 데이터의 중복제거를 통해 데이터의 품질을 향상 시킨다.

○ NTIS 표준코드 매핑
- 텍스트로 입력된 소속기관명 및 학위취득대학교명의 코드매핑
- 소속기관 및 경력사항 직위명의 코드매핑
- 학과명 및 전공명의 코드매핑
- 자택주소 및 소속기관주소의 코드매핑

○ 가비지 처리
- 모든 정제 대상 항목에서 가비지 데이터를 정제
- 전화번호, E-mail, 우편번호 항목의 형식 오류정정
  · 전화번호, 핸드폰, E-mail, 우편번호, 날짜 등 표준포맷으로 변경
  · 기간(시작일, 종료일) 항목은 종료일보다 시작일이 작아야 한다는 논리적 기준에 맞춰 정정
- 논문, 저역서명, 최종학위논문명의 국·영문 입력 위치 오류 정정

○ 이중 등록데이터 처리
- R&D참여인력(주민등록번호 기준)별 소속기관, 학력, 경력, 실적정보에서 2개 이상 존재하는 동일 데이터를 삭제

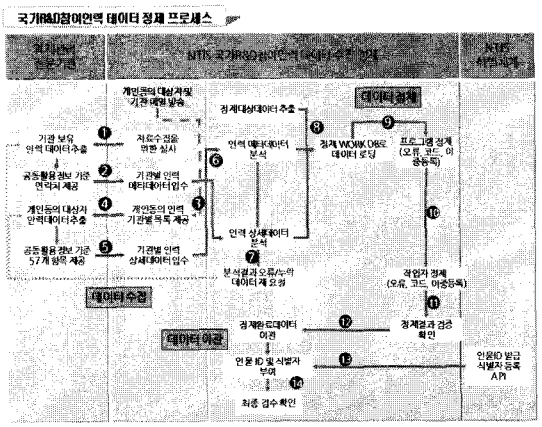

그림 4. R&D참여인력 데이터 정제업무 프로세스
Fig. 4 R&D Project Participants Information Data Refining Process

R&D인력정보 데이터에 대한 단계별 정제업무 프로세스는 다음과 같다.

① 사전 연락 후 직접 방문 및 데이터 입수 관련 협의
② 공동활용정보 기준으로 기관 보유 인력의 연락처 제공
③ 개인동의 협조를 위한 공문 및 안내문 메일링 처리
④ 개인정보 공동활용 동의 대상자 목록 추출 및 해당 기관에 전달
⑤ 공동활용정보 기준으로 기관이 보유한 개인동의 대상 인력의 상세 데이터(57항목) 제공
⑥ 기관제공 인력데이터 현황 및 오류/누락 항목 내역 분석 정리
⑦ 기관제공 인력데이터 분석결과 기관에 통보, 오류 및 누락 데이터 재 수집 요청
⑧ 기관인력데이터 및 수집 구축된 인력데이터를 정제 WORK DB 구조로 전환하여 로딩
⑨ 프로그램 및 SQL을 통한 시스템정제(오류패턴, 코드매핑, 항목별 이중등록 제거)
⑩ 정제작업자에 의한 수작업 정제(오류패턴, 코드매핑, 항목별 이중등록 제거)
⑪ 정제 결과 인력데이터의 검증 확인 및 정제 미비사항 보완
⑫ 정제 완료된 인력데이터를 서비스DB로 이관
⑬ 식별체계에서 제공되는 API를 통한 인물ID 부여 및 식별자 등록
⑭ 최종완료 DB(데이터) 최종 검수 확인 및 이관

4.2 NTIS의 국가R&D과제/성과 데이터 정제

NTIS의 국가R&D과제/성과 데이터는 15개 과제관리기관에서 Bulk(Off-Line) 또는 정보연계(On-Line) 형태로 수집하여, NTIS 국가R&D과제/성과정보 데이터 정제 지침 및 매뉴얼에 따라 프로그램에 의한 기계작업과 정제 작업자에 의한 수작업으로 데이터 정제를 수행하고 있다. <그림 5>는 NTIS의 국가R&D과제/성과 데이터에 대한 데이터 정제업무 프로세스를 나타내고 있다.

데이터의 정제는 NTIS연계시스템으로부터 수집된 국가R&D과제/성과정보 데이터를 NTIS 국가R&D과제/성과정보 데이터 정제지침에 따라 정제업무 프로세스를 수행한다.

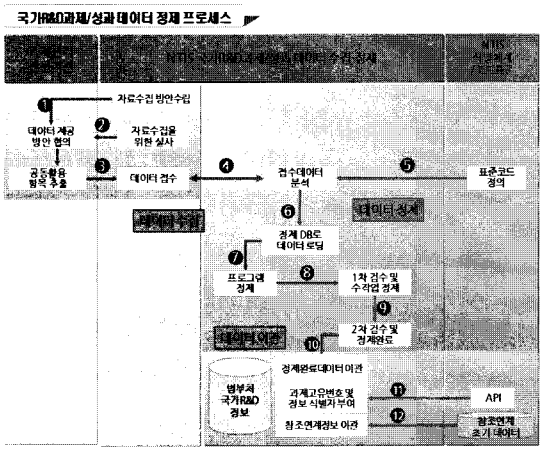

그림 5. R&D과제/성과 데이터 정제업무 프로세스
Fig. 5 R&D Projects/Results Information Data Refining Process

NTIS 국가R&D과제/성과정보 데이터에 대한 단계별 정제업무 프로세스는 다음과 같다.

① 자료수집 방안 수립
② 사전 연락 후 직접 방문 및 데이터 관련 협의
③ 데이터 접수
④ 접수된 데이터에 대한 항목 및 건수 분석
- 데이터 패턴 및 오류패턴 분석 보고서 피드백
⑤ 정의된 표준코드 제공받아 정제대상 데이터의 표준코드매핑 수행
⑥ 원본 데이터 정제DB로 데이터 로딩(중복 데이터 제거)
⑦ 프로그램을 통한 시스템정제(주요항목 체크 및 데이터 오류패턴검출)
⑧ 1차 검수 및 수작업 정제 수행(검출된 오류 항목 정제 및 검수)
⑨ 2차 검수 및 정제완료(검출된 오류 항목 정제 및 검수)
⑩ 정제 완료된 데이터 이관
⑪ 식별체계에서 제공되는 API를 통한 과제고유번호 및 성과정보 식별자 부여
⑫ 과제-성과-인력 참조연계 초기 데이터를 제공받아 이관

V. NTIS 국가 R&D정보의 품질향상을 위한 품질문제 요인분석 및 개선방안

데이터 품질에 문제가 발생할 수 있는 요인은 다양하다. 데이터 품질 문제는 단편적인 이유에서 발생하지 않고, 다방면의 다각적인 이유가 원인이 되어 데이터 품질 문제가 발생하게 된다. 그 중 몇 가지 주요한 요인을 살펴보면 다음과 같다.

첫째 개별 시스템 위주의 개발 및 관리로 인하여 데이터가 분산되고 중복 및 불일치가 발생할 수 있다. 또 데이터간의 복잡한 인터페이스로 인하여 유지보수의 어려움이 가중되어 데이터 품질 문제가 발생할 수 있다.

둘째 원천 데이터의 오류로 인하여 계속적인 정보 품질 저하가 초래되고, 데이터의 속성을 잘못 이해하여 잘못된 용도로 사용하는 경우가 발생할 수 있다. 이로 인하여 결국 데이터에 대한 불신이 쌓이게 되고, 시스템 전체에 대한 불신과 불만으로 이어지게 된다.

셋째 데이터의 변경이나 오류에 대한 수정이 개인적으로 이루어지므로 데이터에 대한 모호성이 증가하여 데이터 품질 문제가 발생할 수 있다. 따라서 체계적인 데이터 관리가 필요하다. 이처럼 데이터 품질문제는 복잡한 원인에서 비롯되기 때문에 체계적이고 지속적인 관리방안을 마련하는 것은 매우 중요하며 고품질의 데이터를 확보하기 위한 데이터 품질관리 활동은 반복적이고 지속적으로 수행되어야 할 것이다.

NTIS는 고품질의 데이터 확보를 위하여 국가R&D 데이터 품질관리체계[15]를 마련하고 데이터의 품질기준과 데이터 품질진단 및 관리 프로세스를 정의 하고 있다. 또 국가R&D 인력/과제/성과정보 데이터 정제 지침 및 매뉴얼[16]을 두어 이를 바탕으로 NTIS 데이터 품질향상 프로세스를 수행하고 있다. 이와 같이 NTIS는 비교적 체계적인 품질관리체계를 갖추고 있다고 할 수 있겠다. 그렇지만 NTIS의 경우 앞서 살펴본 데이터 품질문제 발생의 주요 요인들을 모두 포함하고 있고 할 수 있다. 따라서 NTIS가 내포하고 있는 데이터 품질문제 발생의 원인이 되는 요인들을 분석하고 그 해결방안을 모색하는 것은 NTIS 전체 시스템의 품질향상을 위한 매우 중요한 일이 될 것이다.

지금까지 우리는 NTIS의 국가R&D참여인력/과제/성과정보의 정보연계 구조 및 데이터 정제 방법을 살펴보았다. 이를 통해 본 장에서는 NTIS의 국가R&D정보 연계구조와 데이터 정제 프로세스에 존재하는 데이터 품질문제의 요인을 분석하고 개선방안을 제시한다.

5.1 국가R&D 정보연계 구조의 품질문제 발생요인 분석

본 절에서는 국가R&D참여인력/과제/성과정보의 정보연계 구조에서 품질문제가 발생할 수 있는 요소를 살펴보도록 한다.

첫 번째는 NTIS 표준 스키마와 코드매핑에 따른 품질문제가 발생할 수 있다. 이는 NTIS가 R&D정보에 대하여 기관 레거시DB와 NTIS 통합DB간의 스키마 변경이 여러 번 이루어지고 있기 때문이다. 그 과정에서 각 기관별 상이하거나 특화된 코드를 NTIS에서 모두 수용하지 못함에 따라 품질문제가 발생할 수 있다. 이것은 <그림 2>와 <그림 3>을 통해서 확인할 수 있다. 기관DB에서 정보연계서버의 NTIS공동활용연계DB로 데이터를 넘겨줄 때 NTIS 수집/연계 스키마/코드매핑의 1차 변경이 이루어지고, NTIS 연계 SEED DB에서 NTIS통합DB(국가R&D참여인력/과제/성과정보 포함)로 넘겨질 때 2차 스키마 변경이 이루어진다. 2차 스키마변경 때 NTIS의 표준스키마와/표준코드매핑이 이루어지게 된다. 이와같이 레거시DB에서 통합DB까지 잦은 스키마 변경과 코드매핑은 데이터 품질저하의 원인이 될 수 있을 것이다. 이를 확인하기 위해 데이터 품질분석용 소프트웨어인 DQMiner를 이용하여 관계 및 코드 분석을 실시하였다. 관계 및 코드 분석은 참조관계 또는 코드 일관성 결함으로 발생된 비일관된 데이터를 발견하는 분석 기법이다. 일반적으로 통칭하여 데이터 구조 분석으로 일컫는다. 데이터 구조 분석은 잘못된 데이터 구조로 인해 데이터 값에서 일관되지 못하거나 부정확한 값을 파악하는 분석 기법이다. 또한 데이터 구조적 완전성(누락) 문제로 인해 데이터의 일관성이 결여되는 데이터 값을 발견하고, 사전 정의된 구조 외의 누락된 구조를 발견하고 테스트하여 정확한 구조를 파악하는 것을 주된 목적으로 한다. 분석 대상은 NTIS 연계 SEED DB의 각 과제관리기관별로 저장된 국가R&D참여인력/과제정보 데이터로 하고 ERD분석, 표준 관리 문서 분석을 통하여 얻어진 업무규칙을 기준으로 대상코드에 대하여 관계분석과 코드분석을 각각 실시하였다. NTIS의 경우 관계분석을 하였으나 데이터베이스 관계무결성을 만족하고 있었기 때문에 관계오류는 없었다. 하지만 코드분석의 경우 여러

부분에서 오류가 많이 나타나고 있었다. <표 1>은 NTIS 연계 SEED DB에 있는 데이터 중 현재 분석 가능한 12개 기관의 데이터에 대하여 코드유효성 및 테이블간 일관성을 분석한 결과를 나타내고 있다.

표 1. 관계 및 코드 분석 결과
Table. 1 Relation and Code Analysis Results

구분	대상코드	분석지표	점검 건수	오류 추정 건수	오류 추정율 (%)
인력	표준코드 (국가코드)	코드유효성	23740	3896	16.41
과제	표준코드(학위)	테이블간 일관성	64370	42797	66.49
과제	표준코드 (참여연구기관)	코드유효성	48228	11174	23.17
과제	전공코드	코드유효성	3176	2	0.06
과제	국가기술 지도코드	코드유효성	826	188	22.76

코드 분석을 실시한 결과 오류추정율은 적게는 0.06%~23.17%의 오류가 있는 것으로 나타나고, 표준코드(학위)의 경우는 각각 66.49%로 오류추정율이 비교적 높은 것으로 분석되었다. 이것은 각 기관마다 다른 코드 체계와 표준코드들을 보유하고 있기 때문이고, 수집과정에서 이를 통합표준코드로 매핑하거나 변환하는 과정이 미흡하기 때문에 발생한 한 것으로 보인다.

두 번째는 NTIS의 정보연계 과정 중 정보를 입력하는 창구가 다양하여 품질문제가 발생할 수 있다. 이것은 <그림 2>와 <그림 3>을 통해서 확인할 수 있다. NTIS의 국가R&D참여인력정보의 입력 창구를 살펴보면 기관 레거시DB와 NTIS공동활용연계DB의 연계를 통한 정보입력, NTIS 통합 홈페이지를 통한 NTIS 통합DB로의 정보입력, 기관의 정보연계서버의 SIMS DB를 통한 NTIS 통합DB로의 정보입력, 그리고 정보연계를 통하지 않고 오프라인으로 수집된 Bulk Data를 일괄배치를 통하여 NTIS 통합DB로의 정보입력 등이 있다. 이렇게 다양한 입력 창구를 통하여 정보가 입력됨에 따라 중복 데이터가 발생할 수 있고, 오류 데이터도 많이 발생할 수 있을 것이다. 이것은 결국 NTIS 전체 시스템의 데이터 품질저하의 원인이 될 수 있을 것이다.

세 번째는 국가R&D참여인력 정보연계 프로세스에서 인력 데이터의 주기적(none-Realtime) 전송으로 인하여 품질문제가 발생할 수 있다. 국가R&D참여인력정보의 특성상 특정 과제기관에서 변경/수정된 인력정보는 그 즉시 NTIS 통합DB와 해당 인력정보를 보유하고 있는 다른 과제관리기관에도 즉시 반영되어야만 한다. 그렇지 않는다면 여러 과제 관리기관에서 변경/수정된 정보의 최신성을 파악할 수 없게 되고 결국 최근에 변경된 정보보다 이전에 변경된 정보가 최신정보로 입력되는 문제가 발생하게 된다.

5.2 국가R&D 정보연계 구조의 품질문제 개선방안

앞서 살펴본 국가R&D참여인력/과제/성과정보 연계구조에서 품질문제가 발생할 수 있는 요인을 정리 해보면 크게 두 가지로 요약할 수 있다. 그 첫 번째는 데이터 통합에 있어서 데이터의 표준화와 데이터 코드매핑에 따른 오류가 문제가 됨을 알 수 있고, 두 번째는 국가R&D참여인력정보의 정보연계 프로세스에서 데이터의 주기적(None-Realtime)전송에 따른 데이터 불일치 및 최신성 오류가 문제가 됨을 알 수 있다. 이것은 동일인이 여러 과제관리기관의 과제를 수행하였거나 수행하고 있을 때 수정사항이 발생하면 해당 과제관리기관의 웹 사이트를 방문하여 자료를 수정해야 하므로 동일인에 대한 각 기관들의 자료가 일관성이 없을 수 있다 [17].

5.2.1 데이터 표준화와 데이터 코드매핑에 따른 오류문제 해결을 위한 개선방안

본 절에서는 5.1절의 분석에서 도출된 첫 번째 문제에 대한 개선방안을 논의하고자 한다. 이를 위해 NTIS의 국가R&D참여인력/과제/성과 데이터의 품질진단 결과를 바탕으로 데이터 오류의 원인별 유형을 분석하고, 이를 해결할 수 있는 방안을 제시하도록 한다.

데이터 표준화와 데이터 코드매핑에 대한 데이터의 품질진단 결과를 분석해보면 입력 데이터 품질 통제 미흡, 데이터 요구 증대에 따른 모델 변경, 컬럼 표준화 미흡, 코드 관리 체계 미흡, 데이터 모델관리 미흡으로 크게 5가지로 요약할 수 있다.

▨ 입력 데이터 품질 통제 미흡

- 오류사례
  - FORMAT를 따르지 않는 값을 임의 허용함
  - 날짜 유형의 경우 현재시점과 비교를 실시하지 않음

- 제목, 한글키워드, 영문키워드 등의 컬럼에 한글이 아니거나, 영문문자가 포함되지 않은 문자열 존재
- 원인분석
- 데이터 입력 또는 로딩 시점에 데이터의 유효성 체크 및 검증작업을 실시하지 않음
- 데이터 입력 또는 로딩 시점 또는 사후에 데이터가 Format에 맞는지 재검증하지 않음
- 데이터 입수가 대량으로 이루어질 경우, 다단계 오류데이터 검증 절차가 미비함

▨ 데이터 요구 증대에 따른 모델 변경

- 원인분석
- 과제 데이터의 경우 년차별로 필수 입력 데이터 항목이 조사·분석 기준년도별로 추가되면서, 필수 입력 항목이 증가
- 모델의 잦은 변경과, 복잡도가 심화됨에 따라 정확한 통계데이터를 생성하고 관리하기 어려움

▨ 컬럼 표준화 미흡

- 오류사례
- 동일한 내용인데 상이한 자료유형(Data Type)을 사용
- 데이터 유형 길이가 서로 상이하여 불명확함
- 원인분석
- 각기 다른 개발자 또는 개발시점의 차이로 데이터의 Type, Length등이 상이하게 개발됨
- 데이터 표준을 준수하지 않고 임의 작성하여 의미전달이 정상적으로 이루어지지 않는 경우가 발생함

▨ 코드 관리 체계 미흡

- 오류사례
- 키 컬럼인데 무의미한 값이 입력되어 있거나, 일관되지 않고 두 가지 패턴의 컬럼유형이 존재
- 동일한 코드에 대해 대소문자를 혼용하여 사용함
- 원인분석
- 코드화 수준이 매우 좋으나, 일부 응용프로그램에서 표준을 준수하지 않는 값의 입력을 허용하고 있음
- 코드 정의 및 준수에 대한 기준의 보강이 필요함

▨ 데이터 모델 관리 미흡

- 원인분석
- 논리 및 물리 모델을 관리하고 있으나, ERD상에 엔티티간의 관계가 누락된 사항이 보이고 있음
- 데이터간의 참조무결성을 애플리케이션에서 처리할 경우, 애플리케이션과 DB간의 불일치, 트랜잭션 오류 등으로 인한 중복된 데이터나, 비일관된 데이터가 발생할 가능성이 높으므로 가급적 모델상에 관계를 명시하고, DBMS에 PK, FK로 적용하는 것이 바람직함

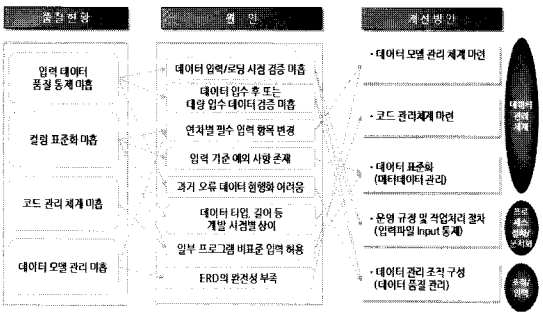

그림 6. 오류유형별 품질문제 해결을 위한 개선방안 제시 모델

Fig. 6 Improvement plan proposed Model for Quality problem solving by Error Type Classification

<그림 6>은 오류유형별 품질문제 원인을 분석하여 그에 적합한 개선방안 제시한 모델을 나타낸다. 이를 종합해 보면 품질 저하 원인으로는 데이터 모델 관리 미흡, 입력 데이터 품질 통제 미흡, 코드 관리 체계 미흡, 데이터 표준화 준수 미흡 등으로 요약 된다.

데이터의 품질의 개선 및 지속적 관리를 위해서는 데이터 관리체계의 수립, 품질 관리 지침 마련, 데이터 표준관리 등이 적절히 수행되어야 할 것이다.

5.2.2 국가R&D참여인력정보의 정보연계 프로세스의 개선방안

본 절에서는 5.1절의 분석에서 도출된 두 번째, 세 번째 문제에 대한 개선방안을 논의하고자한다. 이를 위해 국가R&D참여인력정보의 정보연계 프로세스에서 주기적(Non-Realtime)전송에 따른 데이터 불일치 문제에 대한 해결방안을 제시하도록 한다.

<그림 7>는 NTIS 국가R&D참여인력정보의 정보연계에 있어서 정보연계Agent를 이용한 실시간 데이터 연계 프로세스 방식을 나타내고 있다.

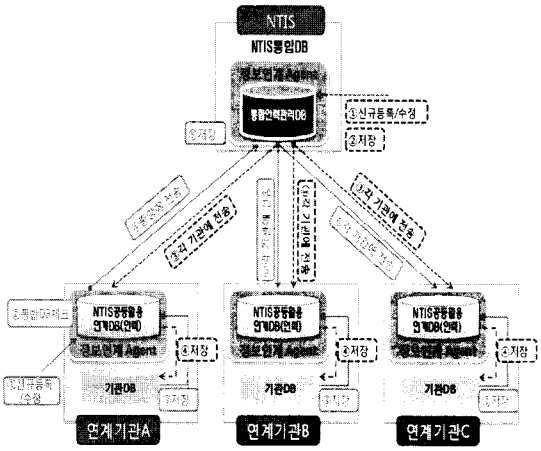

그림 7. 정보연계 Agent를 이용한 실시간 데이터 연계 프로세스
Fig. 7 Real-time Data Linking Process by using Information Linking Agent

정보연계 Agent는 NTIS 시스템에 위치할 중앙 Agent와 각 과제관리기관(연계기관)에 위치할 로컬 Agent로 구성되며, 국가R&D참여인력정보 수집 및 데이터 동기화에 대한 기능을 수행한다. 또한 특정 과제 관리 시스템에서 변경된 인력 표준 항목을 다른 과제관리기관에 적용시킴으로써 특정 인력에 대한 데이터를 보유한 모든 DB의 동기화를 실시간으로 수행하도록 설계하였다. Agent들의 기능은 다음과 같다.

▨ 중앙 Agent
- 로컬 에이전트와 연결 및 요청처리
- 로컬 에이전트로부터 암호화된 XML 문서를 전송받아 복호화한 후 파싱하여 인력정보 추출
- 중복처리 및 오류 검사를 통한 통합인력DB로의 적용
- 매핑정보를 이용한 추출된 인력정보의 변환
- 로컬 에이전트로의 데이터 전송을 위한 XML 생성 및 전송
- 로컬 에이전트 동작 상태 모니터링 및 동작 제어
- 전송데이터 이력 기록 및 조회
- 통합대상기관의 인력DB 스키마 및 XML DTD 관리
- 인력정보의 색인처리

▨ 로컬 Agent
- 중앙 에이전트 연결 및 요청에 대한 동작제어
- 중앙 에이전트로부터 암호화된 XML 문서를 전송받아 복호화한 후 파싱하여 인력정보 추출
- 과제관리기관에서 변경된 인력정보의 XML 형태 변환
- 추출된 인력정보를 통합DB에 전송 및 로그기록

표 2. 인력정보의 신규등록/변경에 따른 통합DB와 연계기관과의 데이터 연계 프로세스
Table. 2 Data Linking Process between Integrated DB and Organizations of Project Management

- 각 연계 기관에서 데이터 수정(①) → 통합DB 복사본 체크하여 데이터를 가져옴(②) → 기관 DB에 저장 및 통합DB복사본에 적용(③) → 통합DB에 전송(④) → 통합DB에 저장(⑤) → 각 기관 통합DB복사본에 적용(⑥)
- 통합DB에 데이터 수정(①) → 통합DB에 저장(②) → 각 기관 통합DB복사본에 적용(③)
- 각 기관에서 신규등록(①) → 통합DB복사본 체크하여 존재하면 데이터를 가져옴(②) → 기관DBdp 저장 및 통합DB복사본에 적용(③) → 통합DB에 전송(④) → 통합DB에 저장(⑤) → 각 기관 통합DB복사본에 적용(⑥)
- 각 기관에서 신규등록(①) → 통합DB복사본 체크하여 미존재하면 기관DB에 신규저장 및 통합DB복사본에 적용(②③) → 통합DB에 전송(④) → 통합DB에 저장(⑤) → 각 기관 통합DB복사본에 적용(⑥)
- 통합DB에 신규등록(①) → 통합DB에 저장(②) → 각 기관 통합DB복사본에 적용(③)

<표 2>은 NTIS 국가R&D참여인력 정보의 실시간 정보연계를 위해 제안된 정보연계 Agent 기반의 정보연계 방식으로 데이터 동기화시 발생가능한 모든 데이터 연계 프로세스를 설명하고 있다.

우리가 제안한 방식을 통하여 모든 과제관리기관에 등록된 국가R&D참여인력 정보는 실시간으로 연계가 이루어지게 되고 인력정보의 중복입력 및 다중수정의 문제가 해결된다. 그리고 언제 어디서 수정 및 변경이 일어나더라도 항상 최신성을 유지할 수 있게 되어 여러 기관에 분산되어 관리되는 정보의 일관성을 유지하게 된다.

5.3 국가R&D정보 데이터 정제 프로세스의 품질문제 발생요인 분석 및 개선방안

NTIS의 국가R&D정보의 데이터 정제 프로세스는 데이터 수집/분석 → 데이터 피드백/재수집 → 데이터 정제 → 데이터 이관의 단계를 거친다. 이 과정에서 데이터

정제는 프로그램에 의한 시스템 정제와 정제작업자를 통한 수작업 정제를 포함하고 있다. 또, 데이터의 수집과정에서 1차적으로 부적합한 데이터를 분석하여 기관에 피드백 함으로써 좀 더 좋은 품질의 데이터를 수집하는 단계를 거치고 있다. 이와 같이 NTIS의 데이터 정제 프로세스 자체는 체계적으로 이루어지고 있음을 알 수 있다. 그렇지만 이러한 정제 프로세스를 통과한 데이터의 품질은 실제 그 과정에서 데이터를 어떤 기준으로 정제하였는지가 매우 중요할 것이다. 따라서 NTIS에 적용된 표준 데이터 정제 룰과 표준 데이터 오류유형 지표가 NTIS의 전체 데이터 정제과정에 있어서 매우 중요한 데이터 품질문제 발생의 요인이 될 것이다. 그러므로 NTIS에 적용된 표준 데이터 정제 룰과 표준 데이터 오류유형에 대한 정확성, 신뢰성을 검증할 수 있는 방안이 필요하고, 지속적으로 새로운 오류유형들을 추가하여 이를 개선해나가는 노력들이 필요할 것이다.

VI. 결론

본 논문에서는 국가과학기술종합정보시스템(NITS)의 국가R&D참여인력/과제/성과정보의 정보연계 방식과 데이터 정제 프로세스를 살펴보고 데이터 품질 저하의 요인을 분석해 보았다. 그리고 정보연계 방식 및 과정에서 발생할 수 있는 데이터 품질문제의 요인을 분석하고 이를 해결하기 위한 개선방안을 제시하였다.

특별히 데이터 통합 및 연계에 의한 스키마 변경과 코드매핑에 따른 품질문제를 데이터 품질진단을 통하여 분석하였고 그 결과로 오류 원인별 유형을 분석하여 개선방안을 제시하였다. 또한 국가 R&D참여인력의 경우 연계방식에 있어서의 주기적(Non-Realtime)연계에 따른 문제점을 지적하고, 이를 해결할 수 있는 정보연계 Agent 기반의 실시간 정보연계 방식을 제안하였다. 그리고 데이터 정제 프로세스의 분석을 통하여 표준 데이터 정제 룰과 표준 데이터 오류유형 지표를 정의 하는 것은 매우 중요한 일이며, 지속적으로 데이터를 분석하는 노력이 필요함을 알 수 있었다.

본 연구를 통하여 여러 기관에 분산되어 관리되고 있는 국가R&D 정보를 통합하고 연계 · 관리하는 것은 많은 문제점을 발생시킬 수 있다는 것을 알게 되었다. 그리고 NTIS의 경우 수집되는 정보의 신뢰성 검증이 취약하다는 점과 정보에 대한 접근 권한 및 복잡한 보안관리, 부처별 타 시스템과의 정보공유, 정보의 중복 및 데이터 요소의 의미적 · 구조적 불일치, 데이터 통합시 시스템간의 이질성, 과제 번호 관리체계의 상이성, 부처간 표준분류체계의 개별성 등의 문제는 앞으로 NTIS가 지속적으로 해결해야할 과제임을 다시 확인 할 수 있었다.

현재 NTIS는 과학기술표준분류 체계를 기반으로 분류체계를 표준화해서 사용하고 있다. 하지만, 각 과제관리기관에서는 각 기관에 특화된 분류체계를 가지고 있기 때문에 이를 모두 수용하기란 현실적으로 어려움이 있다. 하지만, 데이터 표준화, 항목, 코드, 분류체계의 표준화는 지속적으로 이루어져야 할 사항이다. 향후 NTIS에서는 기관별로 특화된 자체분류체계와 과학기술표준분류체계를 모두 포함할 수 있는 표준분류체계를 확립해야 할 것이다.

향후 연구로는 본 연구를 기반으로 하여 NTIS 시스템 전반에 걸쳐 데이터 품질저하 요인들을 체계적으로 분석하고 개선방안을 도출하여 이를 바탕으로 NTIS 시스템에 가장 적합한 품질관리체계를 마련할 계획이다.

참고문헌

[1] 김재수, "국가 R&D 정보 지식포털 NTIS", TLS (Techno Leaders' Digest), 제 219호, pp. 08, 2008.10.

[2] 권도훈, 박성공, 이정욱, "이기종 데이터베이스 환경의 정보 통합을 위한 I2System(:Information Integration System) 설계", Proceedings of the 28th KISS Fall Conference, Oct. 2001 ,pp.136-138

[3] Maurizio Panti, luca Spalazzi, and Allberto Giretti, "A Case-Based Approach to Information Integration", Proceedings of the 26th VLDM Conference, Cairo, Egypt, 2000.

[4] S.Ram, "Special issue on heterogeneous distributed database systems", IEEE Computer Magazine, 24(12), Dec. 1991.

[5] A. Sheth and J. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases",ACM Transaction on Database Systems, 22(3), 1990.

[6] T.Ozsu, U.Dayal, and P.Valduriez, "Distriguted Object Management", Mrgan Kaufmann, San Mateo, CA, 1993.
[7] S.K.Oh, J.H.Son, Y.J.Lee and M.H.Kim, "Efficient Allocation of Workflow Tasks to Improve the Performance of Distributed Workflows", The Fifth Int'l Conference on Computer Science and Informatics, Serial. 5, Atlantic City, USA, P. 445 - 448, 2000
[8] Larry P. English, "Improving Data Warehouse and Business Information Quality", Wiley, 1999.02.
[9] ISO/IEC 9126-1,2,3, JTC1 SC7 WG6(Evaluation & Metrics) Documents, 1996.11.
[10] Dian M. Strong, Yang W. Lee and Richard Y. Wang, "Data Quality in Context", Communications of the ACM, Vol.40 No.5, 1997.05.
[11] 양자영, 최병주, "데이터품질 측정도구", 한국정보과학회 논문지, 컴퓨팅의 실제 제 9권 제 3호, 2003.06.
[12] 한국데이터베이스진흥센터, "데이터품질관리 지침(Ver.2.0)", 2005.11.
[13] 허정희, "데이터 품질 향상을 위한 데이터 관리 프로세스 개선 사례 연구:데이터 표준과 요구사항 관리 중심으로", 한양대학교 석사학위논문, 2007.
[14] 황희정, "데이터 정제 방법론에 관한 연구", 배화논총, Vol.23 No.-, 2004.
[15] 한국과학기술정보연구원, "국가R&D 데이터 품질 관리체계 구축 방안", 2007.
[16] 한국과학기술정보연구원, "국가R&D 인력/과제/성과정보 데이터 정제 지침 및 매뉴얼 Ver 1.0", 2008.
[17] 손강렬, 한희준, 임종태, "산재된 인력정보의 중복 입력 문제 해결을 위한 에이전트 설계 및 구현 방법에 관한 연구", 정보관리연구지 제 38권 제 1호, 2007.03.

저자약력

손강렬(Kang-Ryul Shon)

1999 국립공주대학교
전자계산학과 (공학석사)
2002 국립공주대학교
컴퓨터공학과 (공학박사수료)

2001~현재 한국과학기술정보연구원 책임연구원
※관심분야 : 에이전트, 정보관리, 데이터 통합, 인력정보 시스템