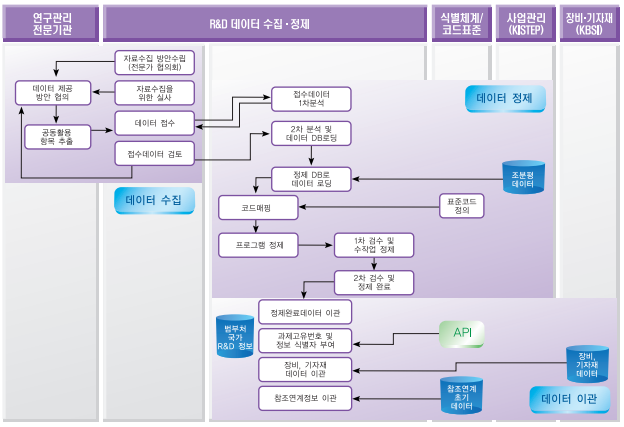
04 국가R&D정보 통합 데이터베이스 구축

글 _ **신성호** 선임연구원, **조항석** 연구원, **손강렬** 팀장 · R&D인력정보팀

1. 개요

국가R&D정보란 범부처적으로 추진되고 있는 R&D사업 또는 과제들의 수행 전후로 발생되는 과제정보, 성과정보, 인력정보, 장비정보 등을 의미한다. 구체적으로는 과학기술관계장관회의(26회, 2007.08.02)에서 확정된 341개 정보 항목으로 구성된다. 이렇게 부처별 · 기관별로 개별 관리되고 있는 국가R&D 관련 정보를 공유 · 공동활용함으로써 국가R&D투자 효율성을 제고시킬 수 있다.

국가R&D 관련 정보를 공동활용하기 위해서는 국가R&D정보의 통합 데이터베이스 구축이 필요하다. 통합 데이터베이스 구축을 위한 기본 프로세스는 데이터 수집, 데이터 정제, 데이터 이관 세 단계로 이루어진다. '07년에는 12개 기관의 데이터를 Bulk(Off-Line) 또는 정보연계(On-Line) 형태로 수집하였고, 수집된 데이터를 NTIS 데이터 정제 지침 및 매뉴얼에 따라 정제를 수행하였다. 정제된 데이터는 정보연계등록모듈을 통해 정보식별자가 부여되어 국가R&D정보 통합 데이터베이스로 구축이 되었다. 이러한 일련의 과정은 〈그림1〉과 같고, 세부적인 내용은 다음과 같다.



〈그림 1〉 통합 데이터베이스 구축 프로세스

2. 데이터 수집

2.1 초기데이터 수집

국가R&D정보 데이터베이스의 원천소스가 되는 데이터는 각 부처별로 지정한 대표연구관리전문기관들로부터 수집되었다. 각 부처별 대표연구관리전문기관들은 각 부처에서 수행된 국가R&D 관련 정보를 통합 · 관리하고 있다. 국가R&D정보는 과학기술관계장관회의(26회, 2007.08.02)에서 의결되었고, NTIS 사업추진위원회 및 전문가협의회를 통해 범부처 협조 차원에서 NTIS사업으로 제공되었다.

데이터 수집은 기본적으로 각 대표연구관리전문기관들과 정보연계서버를 통해 온라인으로 수집되지만, 온라인 수집 체계가 갖추어지기 전에는 Excel이나 DB dump 등의 형태로 FTP나 CD를 통해 수집되었다.

2.2 정보연계를 통한 온라인 수집

'07년에는 대부분의 정보가 Off-Line으로 수집되었으나, 일부 정보는 NTIS와 각 기관 간 정보 S/W에 의한 정보연계를 통해 On-Line으로 수집되었다. On-Line 수집을 위해 각 대표연구관리전문기관에 정보연계 서버를 설치하였다. 이는 기관 Legacy 시스템의 데이터를 정보연계

서버의 데이터베이스로 이관하면, 정보연계 S/W를 통해 데이터를 추출하여 NTIS 중앙서버로 전송하는 방식이다. '08년부터는 대부분의 정보가 정보연계를 통해 온라인으로 수집될 예정이며, 실시간 및 주기적인 정보 수집이 가능할 것으로 기대된다.

3. 수집 데이터 정제

3.1 데이터 정제 지침 및 매뉴얼

데이터베이스 구축의 궁극적인 목표가 되는 데이터 품질 향상을 위해서는 정확한 데이터의 수집도 중요하지만, 수집된 데이터를 체계적으로 정제하는 것이 무엇보다 중요하다. 수집된 데이터의 올바른 정제를 위해서는 정제 기준이 되는 정제 지침 및 매뉴얼이 필요하다.

각 기관들로부터 수집된 데이터를 분석해보면, 다양한 형태의 오류 유형을 발견할 수 있다. 정제 지침 및 매뉴얼에는 '특정 데이터가 오류이다 아니다에 대한 기준, 오류이면 어떤 유형의 오류인지, 그리고 각 유형별 오류를 어떻게 정정할 것인가' 에 대한 세부적인 지침까지 포함하고 있다.

발생 가능한 오류 유형 및 각 오류에 대한 대략적인 정제 지침은 〈표 1〉과 같다. 정의된 정제지침 및 매뉴얼은 실 데이터가 수집된 후 분석한 결과를 반영하여 수정 · 보완하는 과정을 거쳤다.

〈표 1〉 데이터 오류 유형 및 정제지침

오류유형		오류 예제	정제지침
내용 오류		Not Null 항목이나 NULL 값	– 정제가 불가능하므로 정보 재수집 필요
		Not Null 항목은 아니지만 NULL 값	– 정제 불가능(수집 가능한 정보는 재수집)
		중복 레코드	– 중복 기준을 세우고, 중복된 레코드는 삭제
		사업–과제–성과 연계 오류	– 재수집을 통해 정보 간 연결키값 입수
형식 오류	코드	코드 항목이나 명칭으로 존재	– 정의된 표준코드로 매핑
		코드항목이나 표준코드로 매핑 불가	– 코드 표준화 방안에 의거하여 처리
	날짜	YYYY 형식 오류	– 정규 형식(YYMMDDHHMM, YYMM DD, YYYY 등) 이외의 값은 오류 처리 – 상세한 값이 입력된 경우에는 정규 형식으로 매핑 – 정보가 없는 경우 오류 처리(오류표시, 삭제 대상)
		YYYYMM 형식 오류	
		YYYYMMDD 형식 오류	
		날짜가 아닌 데이터	
	숫자	정수 단위 오류	– 정규 형식으로 변환, 변환이 불가능한 경우 삭제
		소수점 이하 자리수 오류	
		숫자가 아닌 데이터	
	문자	특정문자 반복(Garbage성으로 판단)	– 의미 없는 특정 문자는 표시(ex. ###반복)후 삭제
		특정문장 반복(Garbage성으로 판단)	– 의미 없는 특정 문장은 표시(ex. ###반복)후 삭제
		유니코드, 특수문자, CR-LF 포함	– 코드 변환, 텍스트 항목의 CR-LF는 수정 없음

〈표 2〉 기관별 데이터 수집 현황

구 분	기 관 명		과제정보	성과정보	인력정보	장비정보	성과물
대표 연구관리 전문기관	한국과학재단		14,000	66,700	19,600	–	–
	한국학술진흥재단		38,000	63,400	14,300	–	–
	한국산업기술평가원	산업자원부	20,000	–	58,700	–	–
		중소기업청	17,300	–	71,500	–	–
	정보통신연구진흥원		14,600	3,000	27,000	20,900	–
	한국건설교통기술평가원		2,200	400	9,500	45,000	–
	농림기술관리센터		7,400	12,500	13,800	250	–
	한국해양수산기술진흥원		1,000	1,300	2,800	–	–
	한국환경기술진흥원		3,700	4,100	14,600	–	–
	한국보건산업진흥원		4,000	16,500	31,200	–	–
	한국문화콘텐츠진흥원		2,900	–	3,900	–	–
	농촌진흥청		3,000	5,700	3,200	–	–
총 계			128,100	173,600	270,100	66,150	–
성과물 전담기관	한국생명공학연구원 생물자원센터		–	–	–	–	5,000
	컴퓨터프로그램보호위원회		–	–	–	–	5,400
조분평	한국과학기술기획평가원		149,500	137,830	54,000	–	–

3.2 수집 데이터 분석 결과

수집데이터에 대해 테이블별, 레코드별, 항목(column)별 건수 분석을 하였고, NTIS 표준스키마(공동활용정보항목 포함) 및 표준코드에 적합하게 수집되었는지 분석하였다. 이후 각 세부 항목(column)별 특징이나 오류의 유무를 분석하였다. 여기서는 대략적인 수집 건수 현황과 공동활용정보항목(column)과의 매핑 분석 결과 위주로 기술하였다.

〈표 3〉 정제 단계별 수행 업무

정제단계	단계 설명
프로그램 검증	• 수집데이터를 정제DB로 로딩 • 시스템(프로그램) 정제 대상 선정 및 시스템 정제 • 각 작업 단위 완료 시점에 정제 로그 생성
1차 정제 및 검수	• 시스템(프로그램) 정제 후 수작업이 필요한 부분 선별 • 정제매뉴얼에 의거 1차 정제 수행 • 1차 정제 완료 시 정제작업자가 직접 1차 검수를 실시하고, 해당 데이터를 2차 정제 대상으로 변경 • 각 작업 단위 완료 시점에 정제 로그 생성
2차 정제 및 검수	• 1차 검수 후 2차 정제자가 정제매뉴얼에 따라 2차 정제를 수행 • 2차 정제 완료 시 정제작업자가 직접 2차 검수를 실시하고, 정제완료 상태로 변경 • 로그는 각 작업 단위 완료 기준으로 생성되며, 2차 검수와 동시에 정제완료 로그 생성
최종 검수 및 정제완료	• 정제 기준에 따른 최종 검수(샘플링 검수) • 검수 후 적합한 데이터에 한해서 서비스 대상 데이터로 이관 준비

〈표 4〉 데이터 항목(column) 보완 현황

구 분	과제 항목(column)							
	영문과제명	계속과제여부	과제진행상태	다년도협약 구분	부처자체분류_대/중/소	실용화대상 여부구분	연구개발성격구분	연 차
보완 건수	5,800	19,300	2,900	12,000	15,600	3,300	9,400	8,700

구 분	성과 항목(column)									
	논문						특허			사업화
	초록	국내외 구분	학술지 게재 년월일	학술지 시작 페이지	학술지 끝 페이지	학술지 임팩트 팩터	산업재산권 종류구분	특허_삼극대응 취득여부	특허_PCT 출원여부	사업화 년도
보완 건수	54	3,600	3,600	3,600	3,600	1,100	500	200	90	7

3.2.1 수집 건수 분석

'07년에는 13개 부처청이 지정한 12개 대표연구관리전문기관으로부터 과제 약 128,100건, 성과 약 173,600건, 인력 약 270,000건, 장비 약 66,100건 등 약 64만 건의 정보를 수집하였다. 기관별 수집 건수는 〈표 2〉와 같다. 12개 기관과는 별도로 한국과학기술기획평가원(KISTEP)으로부터 조사·분석·평가시스템(KORDI)의 과제 약 15만 건, 성과 약 14만 건을 수집하였고, 성과물전담기관인 한국생명공학연구원 생물자원센터, 컴퓨터프로그램보호위원회로부터 각각 5천여 건의 성과물 정보를 수집하였다.

3.2.2 공동활용정보 항목과 매핑 분석

12개 대표연구관리전문기관으로부터 수집된 데이터가 모두 서비스 대상이 되는 것이 아니라, 341개 공동활용항목에 해당되는 항목(column)만 서비스 대상이 된다(341개 항목 중 대표연구관리전문기관들로부터 수집 가능한 항목은 249개 항목). 따라서 서비스 대상 항목(column)을 선별하기 위해 수집된 데이터와 공동활용항목을 매핑하는 작업이 필요하다. 매핑 결과, 기관별 평균 100여개 항목을 보유하고 있는 것으로 나타났다.

3.3 데이터 정제

데이터 정제 지침 및 매뉴얼을 기준으로 프로그램 정제, 1차 정제 및 2차 정제의 과정을 거쳐 정제를 수행하였다. 프로그램 정제는 기계적(일괄/배치)으로 수정이 가능한 데이터에 대해 실시되었으며, 텍스트성 항목의 오류와 같이 정제작업자가 일일이 확인해야 하는 경우에는 수작업에 의해 정제를 하였다. 정제작업자의 수작업 정제를 돕기 위해 간단한 정제프로그램을 만들었으며, 정제프로그램은 1, 2차 정제 대상을 구분해주어 정제 대상 식별을 용이하게 하였다. 〈표 3〉은 정제단계별로 수행되는 업무를 설명하고 있다.

정제지침 및 매뉴얼을 기준으로 정제단계에 따른 정제 후에는 조사·분석·평가시스템(KORDI)의 과제 및 성과 정보와 비교하여 동일 과제정보, 동일 성과정보에 대해 항목 보완을 실시하였다. 동일 정보(record)임을 식별하기 위해 각 대표연구관리전문기관으로부터 수집된 정보(record)와 조사·분석·평가시스템(KORDI) 정보(record)의 고유 키값을 비교했다. 조사·분석·평가시스템(KORDI)의 정보(record)도 국가R&D과제를 수행한 기관으로부터 수집한 국가R&D정보이므로 해당 기관에서 관리하는 고유 키값을 어느 정도 포함하고 있다.

두개의 키값을 비교하여 동일 정보라고 판단되면, 정보 간 비어 있는 항목(column)을 보완하여 보다 충실한 정보(record)를 확보할 수 있었다.

과거 정보들('06년 이전)은 각 기관에서 관리하는 고유 키값이 변경될 수도 있고, 조사 · 분석 · 평가 시스템(KORDI)에서 가지고 있는 각 기관의 고유 키값이 손실될 수도 있기 때문에, 키값으로 비교하는 데는 한계가 있었다. 따라서 기계적인 키값 비교를 통해서는 일부분의 항목(column)만 보완이 가능하였다.(표 4) '08년에는 정제작업자에 의한 수작업을 통해 과제(성과)명, 항목명(column), 기준년도, 고유 키값 등 세부적으로 비교하여 항목 보완을 할 계획이다.

3.4 정제 결과

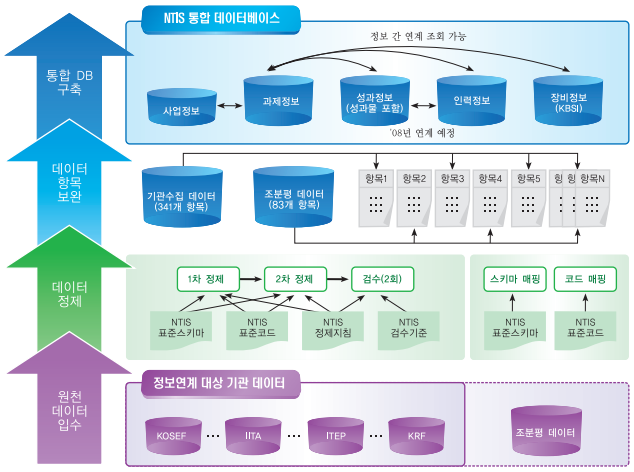
정제지침 및 정제매뉴얼에 의해 정제한 결과는 정제 후 건수(정제 결과)로 나타낼 수 있다. 〈표 5〉와 같이 수집건수가 전체 정제 대상 건수가 된다. 과제 및 성과정보의 경우 데이터 별 고유의 키값을 비교하여 중복된 데이터를 제거하였고, NTIS의 서비스 대상 년도인 '02~'07년을 벗어난 데이터도 역시 배제되었다. 년도가 불분명한 데이터까지 제외되면, 최종적으로 서비스 대상 데이터가 선별된다. 참고로, 성과물전담기관에서 수집된 성과물 정보와 조사 · 분석 · 평가시스템(KORDI)으로부터 수집한 과제/성과정보는 정제된 상태에서 수집되기 때문에 정제 대상이 아니다.

4. 국가R&D정보 통합 데이터베이스 구축

〈그림 2〉는 국가R&D정보 통합 데이터베이스 구축을 도식화 한 것이다. 국가R&D정보 통합 데이터베이스에는 과제정보(사업정보 포함), 성과

〈표 5〉 데이터 정제 결과

구 분	수집건수	중복제거	년도 이외	년도 미상	정제 후 건수 ('02~'07년)
과 제	128,100	1,400	30,600	380	95,720
성 과	173,600	0	26,200	2,300	145,100
총 계	301,700	1,400	56,800	2,680	240,820



〈그림 2〉 통합 데이터베이스 구축 개념도

〈표 5〉 통합 데이터베이스 구축 현황

기준 년도	과제 정보	성과정보						인력 정보	장비 정보	성과물 정보
		논문	특허	기술료	사업화	연수지원	인력양성			
2002	23,100	600	500	2,000	300	9	57	60,300	51,500	10,400
2003	25,900	1,700	1,400	2,300	400	30	170			
2004	26,600	3,100	2,300	2,100	600	24	260			
2005	30,600	34,000	10,000	1,400	200	61	8,500			
2006	32,100	33,700	12,800	1,000	2,500	1,300	6,400			
2007	11,100	6,500	1,500	4	1	300	0			
총계	149,400	79,600	28,500	8,804	4,001	1,724	15,387	60,300	51,500	10,400
		138,016								

※ 성과정보는 당해 연도 이후에 발생 · 입력되기 때문에 '07년 성과정보는 '08년에 본격적으로 수집 · 구축 예정

정보, 인력정보, 장비정보, 성과물정보가 포함되어 있다. 사업정보, 성과정보, 인력정보, 장비정보, 성과물정보는 모두 하나의 과제와 연관된 부수적인 정보들이다. 하나의 과제가 수행되면, 그 과제가 발생된 상위의 사업정보, 과제 수행 결과 발생되는 성과정보 및 성과물정보, 과제에 참여한 참여자 정보(인력정보), 과제에 수반되는 장비정보가 서로 연계되어 사용자들에게 서비스된다.

2008년 3월 현재, NTIS의 국가R&D정보 통합 데이터베이스에는 약 41만 건의 정보가 구축되어 있다(표 6). 구축된 정보들은 과제정보, 성과정보, 인력정보, 장비정보 등 개별적으로 조회가 가능하며, 동시에 식별체계를 부여하여 정보 간 참조 연계가 가능하도록 구축하였다.

5. 향후 추진 방향

이상과 같이 '07년에는 12개 대표연구관리전문기관, 2개 성과물전담기관, 한국과학기술기획평가원(KISTEP)의 조사·분석·평가시스템(KORDI)으로부터 341개 NTIS 공동활용항목을 수집하였다. 수집된 정보는 NTIS 표준스키마 매핑, NTIS 표준코드 매핑 후 정제 과정을 거쳐, 조사·분석·평가시스템(KORDI)의 데이터와 항목 보완을 하였다. 그 결과 약 41만 건의 국가R&D정보 통합 데이터베이스가 구축되었다.

'08년에는 연계대상기관이 확대됨에 따라 신규 데이터가 수집될 예정이다. 이 데이터들도 정제지침 및 매뉴얼에 따라 정제되어 통합 데이터베이스로 구축될 것이다. 정제작업자의 정제 업무를 지원하기 위한 정제시스템을 개발하여 정제업무의 효율을 높이고, 데이터베이스 통합관리시스템 개발을 통해 국가R&D정보를 체계적으로 관리하게 된다. 이밖에 논문 주저자와 과제 참여연구원 연계('06년 데이터), 기관명 DB 구축 등 기 구축된 통합 데이터베이스 보완도 수행될 것이다. 특히 NTIS 공동활용정보 항목 중 기관명이 필요한 항목들에는 공동으로 사용될 수 있는 기관명 DB를 구축하고, 기관 간 공유 및 자동 갱신을 위한 시스템개발이 필요하다. 기관명은 워낙 방대하고, 신규 발생 및 수정이 빈번하게 이루어지기 때문에 다른 코드 항목과는 구별되게 별도의 관리가 필요하고, 연계기관과 상호 공유하여 지속적으로 갱신해 나가야 한다.

구축된 통합 데이터베이스의 품질을 높이기 위한 구체적인 계획도 수립되었다. '07년에는 데이터 품질관리 체계 및 데이터품질관리 시범시스템을 구축하였고, '08년에는 본격적으로 데이터품질관리 시스템을 도입하여 자체적으로 정기적인 데이터 품질 평가를 실시할 예정이다. 품질 평가의 결과를 피드백하여 통합 데이터베이스의 품질 향상을 위해 지속적으로 노력해 나갈 것이다.

참고문헌

[1] 범부처 국가R&D정보 유통기반 구축방안(안), 제26회 과학기술관계장관회의, 2007.8.2
[2] '07년 NTIS사업 보고회 발표 자료, 2007.12.26
[3] 국가과학기술종합정보시스템 구축 사업 최종보고서, 2007.12.31
[4] 관계 부처 실무자급 및 NTIS 전문가협의회 회의 자료, 2008.1.11
[5] '08년 국가과학기술종합정보시스템 구축 사업 사업계획서, 2008.2.13