https://doi.org/10.5392/JKCA.2019.19.10.217

국가R&D정보활용을 위한 전문용어사전 구축

Construction of the Terminology Dictionary for National R&D Information Utilization

김태현, 양명석, 최광남
한국과학기술정보연구원

Tae-Hyun Kim(heemang@kisti.re.kr), Myung-Seok Yang(msyang@kisti.re.kr), Kwang-Nam Choi(knchoi@kisti.re.kr)

요약

국가연구개발(R&D, Research and Development) 정보는 정부부처로부터 발주되는 사업·과제를 수행하는 과정에서 발생되는 다양한 연구분야의 정보들이 포함되어 있다. 따라서 효율적인 R&D정보 검색을 위해서는 이러한 국가R&D정보의 특성을 반영할 수 있는 국가R&D 전문용어사전 구축이 필요하다. 본 연구에서는 국가R&D정보에서 연구분야를 명시하기 위해 활용되는 국가과학기술표준분류를 적용하여 국가R&D용어사전을 구축하기 위한 방안을 제안한다. 국가R&D 과제정보의 구조적 특성과 그에 따른 과제 키워드의 유용성에 대해 언급하고, 국가과학기술표준분류별 국가R&D정보 현황과 국가R&D 용어의 특성에 대해 살펴보고자 한다. 그리고 이를 바탕으로 국가R&D용어사전을 구축하기 위한 방법을 용어사전의 유형과 구조, 사전구축 절차, 정제규칙의 관점에서 정의한다. 본 연구를 기반으로 구축되는 국가R&D용어사전은 국가R&D정보 검색 시 한영 대역어, 동의어 등을 활용한 검색어 확장과 국가과학기술표준분류를 활용한 검색 범위 명확화, 용어설명 정보를 활용한 이용자 편의기능 제공 등에 다양하게 활용될 수 있다.

■ 중심어 : | 국가연구개발정보 | 용어사전 | 과학기술표준분류 | 지능형검색 | 국가과학기술지식정보서비스 |

Abstract

National research and development(R&D) information is information generated in the process of performing R&D based on programs and projects issued by national government departments, and includes information from various research fields as ordered by various departments. Therefore, for efficient R&D information retrieval, it is necessary to build a national R&D terminology dictionary that can reflect the characteristics of such national R&D information. In this study, we propose a method for constructing a national R&D terminology dictionary by applying the classification of science and technology standards used to specify the research field in national R&D information. We will discuss the structural characteristics of national R&D project information and the usefulness of the project keyword, and explain the status of national R&D information by the National Standard Science and Technology Classification(NSSTC) Codes and the characteristics of the national R&D terminologies. Based on this, a method for building a national R&D terminology dictionary is defined in terms of the type and structure of the terminology dictionary, preliminary construction procedures, and refining rules. The national R&D terminology dictionary built on the basis of this study can be used in various ways such as expansion of search terms using Korean-English equivalent words and synonyms when searching national R&D information, clarifying the scope of search using NSSTC, and providing user convenience functions using term explanation information.

■ keyword : | National Research and Development Information | Terminology Dictionary | National Standard Science and Technology Classification Codes | Intelligent Search | National Science and Technology Information Service |

* 본 연구는 한국과학기술정보연구원의「국가과학기술지식정보서비스 사업」으로부터 지원을 받아 수행되었습니다.
* 본 논문은 한국콘텐츠학회 2019 춘계 종합학술대회 우수논문입니다.

접수일자 : 2019년 09월 30일
수정일자 : 2019년 10월 16일
심사완료일 : 2019년 10월 17일
교신저자 : 양명석, e-mail : msyang@kisti.re.kr

I. 서론

최근 인공지능 기술 발전과 함께 지능형 정보서비스 제공을 위한 다양한 노력이 이루어지고 있다. 딥러닝, 데이터 마이닝, 텍스트 분석 등의 빅데이터 분석기법을 적용하여 다양한 분야에 활용하고 있다. 소셜미디어 데이터를 대상으로 한 이용자들의 감성분석을 활용한 마케팅, 뉴스데이터를 이용한 정부정책이나 관련 인물에 대한 긍정·부정도 분석, 연구논문·특허 데이터를 활용한 최신 연구트렌드 분석, 구글이나 네이버 등 포털 이용자의 검색어 트렌드 분석 등 대부분 텍스트를 기반으로 한 분석을 통해 인사이트를 얻어 활용하고 있다. 이러한 텍스트 분석을 위해서는 용어에 대한 정의와 관계 등을 설정하여 분석하는 것이 대단히 중요하다. 특히 과학기술분야나 기타 전문분야에 대한 정보서비스를 위해서는 전문용어 사전에 대한 활용이 매우 중요하다.

국가R&D 정보를 수집·관리하고 서비스하는 국가과학기술 지식정보 포털인 국가과학기술지식정보서비스(이하, NTIS, National Science & Technology Information Service)에서는 17개 부처청으로 부터 국가R&D사업에 대한 과제정보와 논문, 특허 등의 연구성과정보를 수집하여 서비스를 제공하고 있다. 국가R&D사업으로부터 생성된 정보뿐만 아니라 다양한 과학기술정보제공 서비스와의 연계를 통해 과학기술지식 콘텐츠를 제공하고 있다[1][3][4].

최근 NTIS에서는 이용자 맞춤형 검색, 과학기술표준분류추천, 대화형 검색 등 지능형 정보기술을 활용한 다양한 신규서비스를 개발하여 이용자에게 제공할 계획이다[1]. 특히, 인공지능 기술을 적용하기 위해서는 국가R&D정보에 포함된 다양한 전문용어를 이해하고, 분석하는데 필요한 용어사전 구축이 선행되어야 한다. 과학기술 전 분야에 걸쳐 다양한 정보를 담고 있는 NTIS에서는 국가R&D정보의 특성에 맞춰 용어사전을 정의하고 용어 간의 관계정보 등을 활용하여 텍스트 분석을 수행하는 것이 매우 중요하다.

NTIS에서는 그동안 통합검색, 이슈로보는R&D, 과학기술표준분류추천 등의 세부서비스에서 각각의 특성에 맞춰 필요로 하는 용어사전을 별도 관리하고 활용하였다[3]. 이에 따라 국가R&D정보에 대한 용어정의, 관계, 활용방법 등에 대한 표준이 마련되지 않아 NTIS 내부 혹은 다른 서비스에서 활용하는데 어려움이 있었을 뿐만 아니라 일관된 용어처리 결과를 제공하지 못하는 한계가 있었다. 기존 문제점들을 개선하여 검색 효율성을 높이고, 보다 정확한 텍스트 분석결과를 제공하기 위해 국가R&D정보의 특성을 반영한 전문용어사전을 통합·구축할 필요가 있다.

이에 본 연구에서는 국가R&D정보의 중심이 되는 과제정보를 활용하여 국가R&D용어사전을 구축하는 방안을 제안하고자 한다. 국가R&D과제의 협약서를 기준으로 수집되는 과제정보는 연구분야를 명시하는 국가과학기술표준분류와 과제의 핵심단어를 명시하는 한글/영문 키워드를 포함하고 있어 국가R&D의 전문분야정보를 포함하는 용어사전을 구축하기에 용이하다.

우선 본 연구에서는 현재 NTIS가 보유하고 있는 과제정보 내의 과제의 키워드와 국가과학기술표준분류 정보가 용어사전을 구축하기에 적합한지 확인하기 위해 연도별 과제건수를 기준으로 키워드 보유 과제건수, 키워드 중 한글/영문 키워드 개수 일치 건수, 매해 신규로 진행되는 과제건수 등 키워드를 중심으로 한 정보보유현황과 국가과학기술표준분류의 연구분야분류 대·중분류별 과제분포, 대·중·소분류별 평균 과제건수 등을 분석하였다. 그리고 실제 과제 키워드에 입력된 용어들을 상세 분석하여 국가R&D에서 쓰이는 전문용어들의 특이점, 실제 사전 구축 시 고려해야 할 사항들 등을 도출하였다. 이를 토대로 국가R&D 전문용어사전을 구축하기 위해 필요한 용어사전의 유형과 관리구조, 구축절차 등을 정의·설계하고 실제 구축의 각 단계에서 필요한 단계별 용어 정제 규칙 등을 정의하였다.

II. 관련연구

용어사전과 관련한 주요 연구는 1998년부터 2007년까지 10년간 진행된 21세기 세종계획에서의 전문 용어 표준화 사업으로, 3단계로 나누어 연구가 진행되었다. 1단계(1998~2000) 사업에서는 개발 환경 구축 및 기본자료 집성 작업, 2단계(2001~2003) 사업에서는 기 구축된 자원들의 실용화 작업, 3단계(2004~2007) 사

업에서는 집성된 용어 자원들의 효율적인 운용을 목표로 연구가 진행되었다. 표준화용 DB로 약 11만건이 구축되었고, 전문용어 분석/관리/서비스를 위한 소프트웨어 및 웹이 개발되었다[5].

전문용어사전 구축과 관련하여서는 국방과학기술 분야 용어사전 구축[6][7], 정보통신기술 분야 용어사전 구축[8] 등 해당 분야의 특성을 반영한 전문용어사전 구축을 목적으로 다양한 연구가 진행되었다. 그 중에서 국방기술품질원에서 추진한 국방과학기술 전문용어 사전 구축과 관련하여서는 구축과정 자체를 표준화하기 위한 프로세스 표준화와 관련된 연구가 수행되었으며[6], 이러한 표준화를 바탕으로 용어사전을 구축하기 위한 워크밴치를 개발하여 국방과학기술 전문용어사전을 3년 주기로 발간하고 있다[7].

한국정보통신기술협회에서 발간하는 정보통신용어사전은 1993년도 초판 발간을 시작으로 빠르게 변화하고 있는 정보통신기술 분야에 맞춰 새롭게 생성되고 있는 용어들에 대한 표준화된 어휘를 정의함으로써 정보화를 촉진하는데 목적이 있다[8][12]. 해당 연구에서는 외래어로된 용어를 우리말화하고, 잘못 사용되고 있던 용어를 바로 잡는 등 정보통신분야 전반에 활용되는 용어를 재정비하는 작업을 수행하기도 하였다. 매년 신규/갱신 용어를 정리하여 사전에 수록하고, 시사용어집을 발간하는 등 꾸준히 정보통신기술 분야의 용어사전을 구축하여 서비스하고 있다.

정보검색서비스에서 활용 가능한 용어사전을 구축하기 위해 기존 용어사전을 활용하여 확장하는 방법도 제안되었다. 이 방법은 기존 용어사전 자원들을 활용하여 용어사전의 초기 데이터를 쉽게 구조화하고, 신규 발생되는 용어에 대해서는 주기적으로 유사도 매트릭스를 생성하여 활용함으로써 용어사전이 지속적으로 쉽게 갱신될 수 있도록 한 방법이다[9].

III. 연구내용 및 방법

1. 국가R&D 과제정보

NTIS에서 제공되는 국가R&D정보의 핵심이 되는 과제정보는 다음의 [표 1]과 같은 정보 항목으로 구성되어있다[2].

표 1. 국가R&D 과제정보 구성항목

항목명			설명
사업	부처명		연구개발사업의 기획, 평가 및 관리에 관한 제반사항을 주관하는 중앙행정기관의 명칭 예)과학기술정보통신부
	사업명		사업명(세부사업코드의 사업명, 국가연구개발사업 조사분석의 소분류 사업명)
과제고유번호			NTIS에서 발급하는 국가연구개발과제의 범부처 과제고유번호
과제명	국문		신청 및 협약 과제를 기준으로 한 세부과제명의 정식명칭(국문과 영문과제명으로 구분)
	영문		
과제수행기관			연구개발과제를 주관하여 수행하는 기관 명칭
당해연도 연구기간			계속과제의 경우 여러 해에 걸쳐 과제가 진행되는데, 그 중 연구개발계획서나 보고서 등을 작성한 특정 년도에 해당되는 연구기간(시작일, 종료일)을 의미
과학기술 표준분류	연구분야		국가과학기술표준분류 - 연구분야분류 및 가중치, 적용분야분류 및 가중치, 임시분야분류 및 가중치로 구분 - 융합기술인경우 연구분야를 최대 3개까지 입력가능하며, 입력한 각 연구분야에 대한 가중치 (가중치의 합계는 100)
	적용분야		
6T관련기술			6T관련기술코드
연구비	정부연구비		연구비 중 중앙정부의 출연금(원) - 연구개발과제에 배정된 연구비 중 중앙정부에서 투자한 연구비
	민간연구비		연구비 중 민간에서 출연한 금액
...			
요약서	연구목표		개발하고자 하는 기술(공정 또는 제품 포함)의 수준·성능·품질 등 연구목표에 대한 요약
	연구내용		연구내용에 대한 요약
	기대효과		과제 수행 시 기대효과 요약
	키워드	한글	과제 내용을 대표하는 한글 키워드 5개 내외
		영문	과제 내용을 대표하는 영문 키워드 5개 내외

부처명, 사업명 등 발주사업 관련 정보와 NTIS 내에서 과제를 유일하게 식별할 수 있는 과제고유번호, 과제의 국문/영문 명칭, 과제수행기관과 연구기간 등이 있다[2]. 이러한 항목들은 국가R&D과제 수행현황을 기간별, 부처별, 과제수행기관별로 분석하기에 유용한 기본항목이다. 이와 더불어 다양한 부처에서 발주되는 국가R&D과제의 연구분야를 명시·구분하기 위해 과학기술표준분류가 사용되고 있다. 과제에서 사용되는 과학기술표준분류는 연구분야와 적용분야로 나뉘어 기술되며, 융합연구를 표현할 수 있도록 최대 3개의 연구분야, 적용분야 정보를 선택하여 기재할 수 있다. 다건의 분야를 입력하는 경우 가중치를 함께 기재하여 해당 과제가 어떤 분야에 더 연관성이 높은지를 나타낼 수 있다.

각 과제정보에는 연구개발 내용에 대한 요약서를 기재하도록 하고 있다[2]. 이는 연구목표, 연구내용, 기대효과, 키워드로 구성되는데, 과제정보를 검색하고 분석하는데 유용하게 쓰이는 정보이다. 특히 키워드의 경우 해당 과제를 수행한 연구자가 기재한 특정 연구분야에서 쓰이는 전문용어들이 많아 과학기술표준분류의 연구분야 정보와 연계하여 활용하는 경우 검색키워드 확장이나 다양한 지능형 정보서비스에 유용하게 쓰일 수 있다.

표 2. 년도별 과제건수, 키워드 수

기준 년도	총 과제 건수	키워드가 있는 과제(A)		한/영 키워드 개수일치 과제(B)		(B)기준, 총 키워드 개수	(B)기준, 과제당 평균키워드 개수
		건수	비율	건수	비율 (B/A*100)		
2014	52,859	48,367	91.5	43,809	90.6	188,291	4.30
2015	53,719	49,758	92.6	43,364	87.1	179,007	4.13
2016	53,870	50,995	94.7	40,414	79.3	178,347	4.41
2017	60,055	56,245	93.7	46,108	82.0	200,939	4.36
2018	62,271	55,666	89.4	48,912	87.9	217,512	4.45
평균	56,555	52,206	92.4	44,521	85.3	192,819	4.33

NTIS는 '19년 7월 말 기준, 국가R&D로 정부의 연구비 지원을 받아 수행된 과제 76만3천여 건을 서비스하고 있다. 이는 2002년도부터 수집하여 서비스하고 있는 것으로, 최근 5년을 기준으로 볼 때 지속적으로 증가하고 있는 추세이며, 2018년도를 기준으로 볼 때 6만2천여 건의 과제정보가 수집되어 서비스되고 있다[1]. [표 2]에서 보는 바와 같이 평균적으로 약 92%의 과제정보가 키워드를 보유하고 있고, 그 중 한·영키워드의 개수가 일치되는 정보는 약 85.3%에 달한다. 따라서 한·영키워드를 활용해 용어사전을 구축하면, 한글용어에 대한 영문대역어를 함께 구축할 수 있어 대부분 영어로 작성되는 논문성과 정보를 검색할 때 유용하게 사용할 수 있다. 과제 당 평균적으로 입력되는 키워드의 수가 4.3개임을 감안할 때, 연평균 약 19만3천여 건 정도의 유효한 키워드가 등록되고 있다고 볼 수 있다. 2018년도 기준 6만 여건의 과제 중 신규과제 건수는 약 2만2천여 건으로 대략 36.4% 정도의 과제가 새롭게 기획·진행되고 있다. 따라서 매해 새롭게 등록되는 이러한 과제정보들을 활용해 국가R&D용어사전을 구축하는 경우 기 등록된 용어들과의 중복을 고려하더라도 새로운 연구분야에서 발생하는 신조어들을 획득하기 용이하다.

2. 국가과학기술표준분류

국가과학기술표준분류(이하 과기표준분류)는 국가 과학기술 관련 정보·인력·연구개발사업 등의 효율적 관리와 국가연구개발사업의 연구기획·평가 및 관리, 과학기술예측 및 기술수준평가 수행, 과학기술 정보의 관리·유통 등을 위해 과학기술기본법 제27조에 근거하여 2002년도에 제정되었으며, 국가R&D 연구분야의 변화를 반영하여 주기적으로 수정·보완하고 있다. 현재 2018년도 개정 기준으로 연구분야 대분류 33개, 중분류 371개, 소분류 2,898개와 적용분야 대분류 33개로 분류체계가 구성되어 있다[10].

과기표준분류는 범부처적으로 추진되는 국가R&D에 대한 분류를 담기위해 *과학기술* 관련 연구 분야 외에도 *인문사회과학* 분야의 분류까지도 포함하고 있다[10]. 그러나 국가R&D의 특성상 대부분의 과제가 [그림 1]에서 보는 바와 같이 *수학, 물리학* 등을 포함하는 자연분야에서 *기계, 재료, 건설/교통* 등을 포함하는 *인공물* 분야에 이르는 *과학기술* 관련 분야에 집중되어 있다. [그림 1]은 2014년도부터 2018년도 까지 5개 년도에 걸쳐 과기표준분류별 수행과제 건수를 대분류 기준(상단 그래프)과 대분류 내 포함된 중분류 건수를 반영한 기준(하단 그래프)으로 나누어 그 분포를 나타낸 그래프이다. *과학기술* 관련 분야에 해당되는 16개 대분류인 수학에서 *건설/교통*에 이르는 분류에 걸쳐 약 94%의 과제가 집중되어 있고, *인문사회과학* 관련 분야에 해당되는 17개 대분류인 *역사/고고학*에서 *과학기술*과 *인문사회*에 이르는 분류에 걸쳐 6%의 과제가 수행되었음을 알 수 있다.

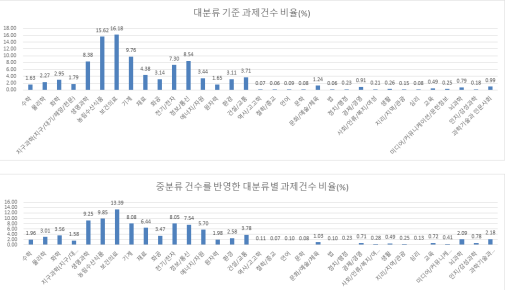


그림 1. 과기표준분류 내 과제건수 비율

다음 [표 3]은 과기표준분류의 대·중·소 분류별 과제 건수를 전체, 과학기술 관련 연구 분야, 인문사회과학 관련 연구 분야로 나누어 본 것이다. 분류의 개수로 볼 때 *과학기술*과 *인문사회과학* 분야의 분류 개수는 중분류 기준으로 보더라도 약 56%, 44%로 크게 차이가 나지 않지만, 과제 건수 기준으로 보면 과학기술 분야의 분류에 약 12배 많은 과제가 있음을 알 수 있다. 이로써 대부분의 국가R&D가 과학기술 분야에 집중되어있음을 알 수 있다.

표 3. 과기표준분류 대·중·소 분류별 과제 건수

	분류 건수			분류별 과제 건수		
	전체	과학기술	인문사회과학	전체	과학기술	인문사회과학
대분류	33	16	17	1,521	2,945	182
중분류	**371**	**208**	**163**	**135**	**227**	**19**
소분류	2,898	1,645	1,253	17	29	2

이와 같은 과기표준분류의 대·중·소 분류 건수와 분류별 과제건수 분포를 기준으로 판단할 때 국가R&D용어사전을 구축하고 활용하는데 있어 의미있는 결과를 얻기 위해서는 용어사전 구축자나 이용자가 쉽게 인지할 수 있는 수준과 기계학습에 활용할 수 있는 적절한 과제건수를 고려할 때 소분류 보다는 중분류 수준에서 국가R&D용어사전의 분류정보를 구축하는 것이 적합하다. 따라서 본 연구에서 이후 언급하는 분류 정보는 과기표준분류의 중분류를 기준으로 한 것이다.

3. 국가R&D 용어 특이점 및 고려사항

국가R&D용어는 동일한 용어일지라도 분류에 따라 다르게 해석되는 특성이 있다. 예를 들어 “Distribution”은 원예작물과학 분야에서는 “보급”, “유통”의 의미로 쓰이지만, IT 분야에서는 “분배”로 해석되어 쓰이고 있으며, “Aging”의 경우 원자력 안전기술 분야에서는 “경년열화”, 기타 과학기술분야에서는 “노화”, “고령화” 등으로 해석되어 쓰인다.

동음이의어의 경우에는 분류정보를 활용하여 의미를 명확화할 수 있다는 장점이 있다. 예를 들어, “배”의 경우 분류정보와 용어의 영문대역어를 활용하면 식량작물과학 분야에서는 “Pear”, 조선/해양시스템 분야에서는 “Ship”, 유전학/유전공학 분야에서는 “Belly”로 주로 쓰이므로, 정보검색 시에 검색결과 집합을 조정하는데 유용하게 쓰일 수 있다. 영문약어로 검색하는 경우에도 분류정보와 결합하면 모호한 의미를 명확화하여 활용할 수 있다. 예를 들어, “AI”의 경우 “인공지능: Artificial Intelligence”, “조류독감: Avian Influenza”, “인공수정: Artificial Insemination”과 같이 연구분야에 따라 전혀 다른 의미의 국문/영문용어로 사용되고 있어 분류정보와 결합하여 검색하면 더욱 정확한 검색결과를 얻을 수 있을 것이다.

전문용어의 경우 외래어표기가 많은데, 이와 관련해 동일 용어를 다양한 외래어로 표기하는 경우 국문키워드만 활용해 정보를 검색하면 기대하는 결과를 얻을 수 없게 된다. 예를 들어, “Apoptosis”는 “아팝토시스”, “아포토시스”, “아폽토시스”, “사포자연사”, “세포사멸” 등의 국문으로 표기되고 있다. 외래어표기와 우리말표기가 혼용되어 사용되고 있으며, 특히 외래어표기의 경우 이외에도 다양한 표기가 존재한다. 따라서 외래어의 경우 가장 많이 사용되는 용어를 대표용어로 정의하고 나머지 용어는 동의어로 정리해 관리할 필요가 있다.

마지막으로 국가R&D 과제정보의 키워드는 과제정보를 등록하는 연구책임자가 개별적으로 기재하는 정보이므로 중복을 고려하더라도 약 4만여 명의 연구책임자가 키워드 정보를 입력한다고 볼 수 있다. 따라서 등록되는 키워드, 특히 영문 키워드에는 다양한 오탈자 및 띄어쓰기 오류가 존재할 수 있고, 단·복수 표기가 혼재한다. 이러한 과제키워드를 활용해 용어사전을 구축하는 경우 어떠한 용어를 표준으로 정의할 지에 대한 정제규칙이 마련되어야한다.

4. 국가R&D용어사전 구축 방안

국가R&D용어사전을 구축하기 위해서는 구축하고자 하는 용어사전의 유형과 구조를 명확화하고, 이에 맞는 용어사전을 구축하기 위한 구축절차와 정제규칙을 정의해야 한다.

4.1 국가R&D용어사전의 유형과 구조

국가R&D1정보를 효과적으로 검색하기 위해서는 연구분야별 전문용어뿐만 아니라, 용어간의 관계 정보, 개체

명, 불용어 등 용어사전의 상세 유형이 식별되어야 한다.

용어관계 정보는 상·하위관계, 포함관계, 단순 관련어를 포함하는 정보로, 별도 사전으로 관리하여도 무방하나, 전문용어의 관련 항목 수준으로 정의하는 것이 검색 및 관리 측면에서 볼 때 보다 효율적이다.

국가R&D정보 내에서 주요하게 식별되어야 하는 개체명으로는 과제수행기관, 과제관리기관 등의 항목에 쓰이는 기관명과 연구책임자, 참여연구원, 평가위원정보 등에서 쓰이는 인명, NTIS 내에서 사용하는 세부서비스, 메뉴, 정보항목 등의 명칭에 해당하는 서비스명, 과제수행기관의 소재지 정보에 해당되는 지명이 있다. 이러한 개체명은 전문용어와는 별개로 구분하여 사전을 구축관리하는 것이 효율적이다. 그밖에 불용어 사전도 별도 구축하여 관리하면 용어사전을 활용하는 서비스들에서 불용어들을 걸러내는데 유용하게 쓸 수 있다.

용어사전을 활용목적에 맞게 다양하게 활용하기 위해서는 용어사전의 구조, 즉 용어에 대한 메타데이터로 어떤 항목들을 둘 것인지를 명확하게 정의하고 구조화하는 것이 중요하다. 본 연구에서 정의한 국가R&D용어사전의 구조는 다음의 [표 4]와 같다.

표 4. 국가R&D용어사전의 구조

속성	설명
용어ID	데이터베이스 내에서 용어를 유일하게 식별하기 위한 ID
국문용어	국문용어
영문용어	국문용어에 대한 영문 풀네임
약어	영문 풀네임에 대한 약어(약어가 존재하는 경우만 기재)
동의어	동의어를 구분자로 구분하여 등록
반의어	반의어를 구분자로 구분하여 등록
상위관계어	현재 용어의 상위 관계에 해당하는 용어의 ID를 구분자로 구분하여 등록
포함관계어	현재 용어를 포함하는 관계에 해당하는 용어의 ID를 구분자로 구분하여 등록
관련어	관련 용어들을 구분자로 구분하여 등록
사전유형	사전의 유형을 구분하는 항목 - 기본용어, 개체명(기관명, 인명, 지명, 서비스명), 비속어
설명	용어에 대한 의미를 설명하는 글
설명출처	용어 설명의 출처 명기
과학기술 표준분류	용어가 속하는 과학기술표준분류 연구분야(중분류 기준으로 관리)를 다건 등록관리(별도 테이블로 정의)
범용여부	용어가 특정 과학기술표준분류에 속하지 않고 범용적으로 쓰이는 용어인지 구분하는 항목
등록일, 등록자	용어를 등록한 날짜와 등록자의 ID
개정일, 개정자	용어를 개정(수정)한 날짜와 개정자의 ID

국영문 용어와 약어를 많이 혼용해서 사용하는 국가R&D용어의 특성 상 *국문/영문/약어* 트리플을 기본으로 하여 용어사전의 구조를 정의한다. *국문/영문* 쌍은 필수 항목으로 정의하며 용어 쌍을 기준으로 중복없이 각 1개 용어를 등록하는 것을 원칙으로 한다. *약어, 동의어, 반의어, 관련어* 등은 경우에 따라 추가될 수 있는데 항목 내에서만 중복되지 않도록 하여 각 항목에 여러 개의 용어를 구분자로 구분하여 등록할 수 있다. *상위관계어*와 *포함관계어*는 현재 용어의 상위 관계에 해당하는 용어, 현재 용어를 포함하는 관계에 있는 용어의 ID를 구분자로 구분하여 여러 건 등록할 수 있다. 이 경우는 단순히 관련 용어 자체를 항목에 등록하는 방식으로는 명확한 상위 또는 포함 관계어를 인식하기 어렵기 때문에 용어ID를 등록하여 관리한다. *사전유형*은 국가R&D용어사전의 기본 목적에 따른 전문용어에 해당되는 경우 구분 값을 “기본용어”로 하여 등록하고, 그 밖의 개체명의 경우 “기관명”, “인명”, “지명”, “서비스명” 등 개체명 유형에 해당하는 구분 값을 등록한다. *설명*에는 해당 전문용어의 의미를 설명하는 글을 등록하고 *설명출처*에 해당 설명글을 참조한 출처를 기재한다. *과학기술표준분류*는 별도의 테이블로 정의하여 등록하며, 해당 용어의 분류정보를 대분류명과 중분류명으로 구분하여 여러 건 등록할 수 있다. 단, 하나의 용어에 대해 대분류/중분류 쌍을 기준으로 중복된 값이 입력되진 않는다. *범용여부*는 용어가 과기표준분류 상의 어떤 분류에도 속하지 않는 범용적으로 사용되는 용어인지 여부를 구분하기 위해 사용되는 항목이다. 그 밖에 데이터 관리 및 기능적인 확장성을 고려하여 *등록일, 등록자, 개정일, 개정자*에 관한 정보를 부가정보로 관리한다.

4.2 용어사전 구축 절차

본 연구에서는 국가R&D 과제의 키워드를 기반으로 국가R&D용어사전을 구축하므로, 키워드의 추출 및 가공으로부터 그 절차가 시작된다.

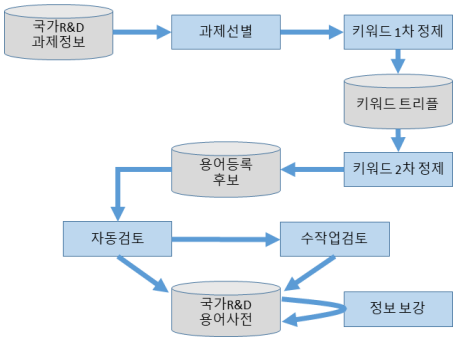


그림 2. 국가R&D용어사전 구축 절차

[그림 2]에서 보는 바와 같이 국가R&D과제정보로부터 우선 키워드가 1개 이상이고 국문/영문 키워드의 개수가 일치하는 과제를 *선별*한다. *키워드 1차 정제*에서는 과제 키워드를 파싱하여 국문/영문 쌍을 만들고 여기에 과기표준분류정보까지 추가하여 키워드 트리플을 만드는데, 이때 키워드의 국문/영문 앞뒤에 들어간 공백이나 하이픈(-), 구두점(.)을 제외한 특수문자를 제거하고, 국문에 영문만 입력되어있다거나 영문에 국문만 입력되어있는 입력 오류데이터는 제외하며, 비교의 효율성을 높이기 위해 국문은 용어 내 띄어쓰기를 모두 배제하고, 영문은 소문자를 모두 대문자로 변환하여 저장한다. 불용어사전에 등록된 키워드의 경우는 제외한다.

*키워드 2차 정제*에서는 키워드 자체에 괄호가 포함된 경우와 영문의 띄어쓰기 오류, 단·복수 표기를 정제한다. 국문에 "인공지능(AI)"와 같이 괄호 안에 영문이 기재된 경우 괄호 안의 영문은 삭제하고, 해당 영문이 트리플 상의 영문과 다른 경우 영문을 구성하는 각 단어의 첫 글자를 조합한 용어와 일치하는지 확인하여 일치하는 경우 약어 항목에 해당 정보를 등록한다. "벼(쌀)"과 같이 국문용어의 괄호 안에 국문이 기재된 경우 괄호 안의 국문은 단순 삭제 처리한다. "5GENERATION(5G)"와 같이 영문에 괄호가 포함된 경우는 괄호 안팎의 용어가 풀네임과 약어의 관계인지 확인하여 풀네임은 영문 항목에 남기고 약어는 약어 컬럼에 별도 등록한다. 영문 키워드 띄어쓰기는 "AIRCONDITIONER", "AIR CONDITIONER"와 같이 띄어쓰기로 인해 다르게 등록된 경우가 있는지 확인하여 띄어쓰기가 포함된 용어로 정정한다. 영문 단·복수 표기의 경우 외부용어사전을 참조하여 의미없는 복수표기는 단수형으로 통일하여 정정한다. "안경/GLASSES"와 같은 의미있는 복수 표기는 유지한다.

*자동검토*에서는 용어등록후보에 있는 용어들의 빈도를 산출하여 고빈도 용어(예: 빈도 20이상)이면서 한글기준 길이가 15자 미만인 용어들을 추출하고, 외부용어사전을 참조하여 정합성을 검증한 후 *국가R&D용어사전*에 자동 등록한다. 이때 정합성 검증에 실패하였거나 국문/영문/약어 트리플 중 1개는 일치하고 나머지가 다른 경우는 *수작업검토* 대상으로 넘긴다. 또한 구두점(.)이나 숫자 등이 포함된 용어, 한글기준 용어길이 15자 이상인 용어, 중빈도 용어(예: 빈도 10이상 20이하)의 경우도 *수작업검토* 대상으로 표시한다.

*수작업검토*에서는 자동검토에서 걸러진 용어들을 대상으로 용어사전검토자가 사례별 정제규칙에 따라 *용어등록후보*들을 검토하여 *국가R&D용어사전*에 용어를 등록한다. 다음의 [표 5]는 수작업검토 시 적용되는 정제규칙의 예이다.

표 5. 국가R&D용어사전 수작업검토 정제규칙 예

케이스	정제 규칙	사례	등록데이터
정관사, 전치사 등이 포함된 영문	정관사, 전치사가 빠져도 의미가 바뀌지 않는 경우 빼고 등록	문화권/ A CULTURAL AREA	문화권/ CULTURAL AREA
	의미가 바뀌는 경우 정관사, 전치사를 유지한 체 등록	지방자치단체/ A LOCAL GOVERNMENT	좌동
		지방자치제/ LOCAL GOVERNMENT	좌동
15자 이상 국문	영문 기준으로 의미 상 문제가 없으면 유지	글리세롤디아세테이트라우레이트/GLYCEROL DIACETATE LAURATE	좌동
	문장형 용어는 삭제	인터액티브하고지능화된이모티콘/INTERACTIVITY INTELLIGENT EMOTICON	삭제
국문 또는 영문 중 하나가 일치되는 경우	영문이 일치 시, 영문이 약어가 아니면 국문 기준으로 빈도가 높은 용어를 남기고, 다른 국문은 동의어에 기재 후 삭제	비생물적스트레스/ ABIOTIC STRESS	환경스트레스/ABIOTIC STRESS/…/비생물적스트레스
		환경스트레스/ ABIOTIC STRESS	
	국문이 일치 시, 의미 상 동의어이면 영문 기준으로 빈도가 높은 용어를 남기고, 다른 영문은 동의어에 기재 후 삭제	간기능/ LIVER FUNCTION	간기능/LIVER FUNCTION/…/HEPATIC FUNCTION
		간기능/ HEPATIC FUNCTION	

정보보강 단계에서는 외부용어사전을 활용하여 설명, 동의어, 반의어 등의 관련 정보를 자동 보강한다. 이를 위해 국립국어원의 표준국어대사전[11], 특허정보 활용 서비스에서 제공하는 특허 시소러스, 영한 특허기술용어 번역사전[12], 한국정보통신기술협회에서 제공하는 정보통신용어사전[13], 국방기술품질원에서 제공하는 국방과학기술용어사전[14] 등을 활용한다.

IV. 결론 및 향후 연구

본 연구에서 제안하는 국가R&D용어사전 구축 방안은 기본용어 구축의 관점에서 상세 기술되었다. 국가R&D과제정보를 기반으로 과제 키워드를 추출하여 국문/영문/과기표준분류의 트리플을 구축하고, 이를 바탕으로 용어를 정제하여 국문용어, 영문용어, 약어, 동의어, 과학기술표준분류 수준까지의 기본정보를 구축한다. 이를 바탕으로 다양한 외부용어사전을 활용해 설명 및 기타 항목을 보강한다.

용어사전을 활용하고자 하는 서비스들의 요구사항은 크게 2가지 유형으로 구분할 수 있다. 전문용어의 연구분야 분류정보를 필요로 하는 경우와 검색어 확장을 목적으로 용어에 대한 동의어와 관련어를 필요로 하는 경우이다. 본 연구에서 제안한 방법으로는 타 용어사전에서 제공할 수 없는 과학기술표준분류 기준의 전문 연구분야 분류 정보를 제공할 수 있다는 장점이 있다. 또한 키워드 빈도를 기준으로 용어를 자동추출하여 용어사전을 구축하므로, 국문/영문 대역어와 함께 동의어까지도 손쉽게 구축할 수 있다는 장점이 있다. 그러나 관련어와 관련해서는 외부용어사전을 활용한 관련어 정보 구축만으로는 신규 용어나 외부용어사전에서 잘 검색되지 않는 전문용어에 대한 관련어 제공 부분이 약할 수밖에 없다. 따라서 향후에는 본 연구에서 제안한 국가R&D용어사전 구축 방법을 확장하여 동일 과제 내에서 공기(Cooccurrence)하는 용어들을 기반으로 용어 관계정보를 산출하여 관련어를 구축하는 방안에 대한 연구가 필요하다.

그밖에도 다양한 외부용어사전들을 효율적으로 참조하고 활용하기 위한 용어사전 표준양식 개발 및 관리 방안에 대한 연구와 용어사전을 보다 효율적으로 구축하고 관리하기 위한 관리도구 개발에 관한 연구도 수행되어야 할 것이다.

참 고 문 헌

[1] 국가과학기술지식정보서비스, http://www.ntis.go.kr

[2] 국가연구개발정보표준 [시행 2018. 9. 27.] [과학기술정보통신부고시 제2018-59호, 2018. 9. 27., 일부개정]

[3] 조우승, 김정오, 박민우, 최기석, 김태현, "A Study on the Construction of the User-Customized Researcher & Research Institute Information Curation System based on National R&D Data," Journal of Advanced Research in Dynamical and Control Systems, Vol.10, No.11, pp.1389-1394, 2018.

[4] 양명석, 강남규, 김태현, 주원균, "Improvement for Generation process of Researchers Map on National R&D data," 2015 International Confernece On Future Information & Communication Engineering, Vol.7, No.1, p.351, 2015.

[5] 홍재성 외, 21세기 세종계획 전자사전 개발 연구 보고서, 문화관광부, 2000, 2001, 2002, 2003, 2004, 2005, 2006.

[6] 최중환, 최석두, 김이겸, 박영욱, 정종희, 안희정, 정한민, 김평, "국방과학기술 전문용어 사전의 구축 프로세스 표준화 및 활용 방안," 한국콘텐츠학회논문지, 제11권, 제8호, pp.247-259, 2011.

[7] 최중환, 박정호, 김경선, 김평, "국방과학기술 전문용어 사전 구축을 위한 프로세스 및 워크벤치 개발," 한국콘텐츠학회논문지, 제12권, 제8호, pp.420-428, 2012.

[8] 문정임, "제4판 정보통신용어사전을 발간하면서…", TTA 저널, 제73호, pp.61-65, 2001.

[9] 최희석, 김무철, 한희준, 김윤정, 김재수, "R&D정보검색서비스를 위한 용어사전 구축 방법," 2013 한국정보과학회 추계학술발표회, pp.513-515, 2013.

[10] 과학기술정책지원서비스 국가과학기술표준분류체계, https://www.k2base.re.kr/clInfo/aboutClInfo.do

[11] 국립국어원 표준국어대사전, https://stdict.korea

n.go.kr
[12] 특허정보 활용서비스, http://plus.kipris.or.kr/portal/main.do
[13] 한국정보통신기술협회 정보통신용어사전, http://terms.tta.or.kr
[14] 국방기술품질원 국방과학기술용어사전, http://www.dtaq.re.kr/ko/state/data_open.jsp

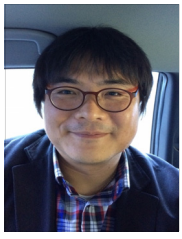
저 자 소 개

김 태 현(Tae-Hyun Kim) 정회원

- 2001년 2월 : 충남대학교 컴퓨터과학과(이학석사)
- 2001년 3월 ~ 2001년 11월 : ㈜엔퀘스트테크놀러지 연구원
- 2002년 3월 ~ 2004년 2월 : 한국전자통신연구원 연구원
- 2004년 3월 ~ 현재 : 한국과학기술정보연구원 선임연구원 / NTIS개발팀장

〈관심분야〉 : 정보검색, 정보분석, 전문용어사전구축, 소프트웨어공학

양 명 석(Myung-Seok Yang) 정회원



- 2001년 2월 : 충남대학교 컴퓨터과학과(이학석사)
- 2017년 2월 : 충남대학교 컴퓨터공학과 박사
- 2001년 3월 ~ 현재 : 한국과학기술정보연구원 책임연구원 / NTIS기획팀장

〈관심분야〉 : 정보검색, 데이터베이스, 네트워크분석

최 광 남(Kwang-Nam Choi) 정회원

- 1994년 2월 : 충남대학교 컴퓨터공학과(공학석사)
- 2017년 2월 : 배재대학교 컴퓨터공학과(공학박사)
- 1994년 7월 ~ 현재 : 한국과학기술정보연구원 책임연구원 / NTIS센터장

〈관심분야〉 : 정보검색, 정보분석, 빅데이터