국립국어원 2020-01-21

발 간 등 록 번 호
11-1371028-000828-01

공공용어 관리 체계 구축을 위한 중장기 계획 수립

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구 용역 계약에 따라 ‘어려운 공공용어 분석 및 개선 지원 시스템 개발’ 사업의 중장기 계획 수립에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2020년 4월 ~ 2020년 12월

2020년 12월 12일

사업 책임자: 이영현(㈜엔에이치엔다이퀘스트)

[차 례]

[표 차례]

[그림 차례]

I. 서론

1. 공공용어 관리 체계 구축 추진 배경 및 현황

전문용어는 각 전문 분야에서 사용되는 단어들을 말한다. 예를 들어 의사들이 병원 등에서 사용하는 용어들은 의학 용어라는 전문용어에 해당한다. 학계에서 사용하는 학술용어, 과학기술 분야의 과학기술 용어, 산업계의 산업용어, 정부업무 분야의 정책용어 및 행정용어 등이 전문용어에 속한다. 전문용어는 '공공언어로서의 전문용어'와 '학술적 전문용어'로 구분할 수 있다. 공공언어로서의 전문용어란 학술적 전문용어에서 출발하였지만, 일반 언중에게 고시되거나 사용될 가능성이 있는 것으로 대중성이 있는 용어를 말한다.

공공 정보 공개의 일상화, 인터넷 발달 등에 따라 지식이 공유되고 확산하면서 '심장병', '인터넷', '콤플렉스'처럼 전문용어인지 일상용어인지 구분이 힘든 용어가 많아지고 있으며, 전문가는 물론 일반인도 전문용어를 많이 사용하고 있다. 문제는 외래어로 된 낯설고 어려운 전문용어 유입으로 정보 소외층이 양산되고 있으며 과학기술 분야 간 전문용어 사용 혼란이 발생한다는 것이다. 이러한 문제점을 해결하기 위해 전문용어 표준화가 필요하다.

전문용어 표준화는 공공성을 띤 전문용어를 국민이 알기 쉽게 다듬고 체계적으로 통일하는 것이다. 표준화 방식을 구체적으로 살펴보면 하나의 개념이 여러 가지 다른 용어로 사용되는 것을 하나로 통일하는 표준화, 어려운 용어를 쉬운 말로 바꾸는 표준화, 어문 규범(한글 맞춤법, 표준어 규정, 외래어 표기법 등)에 어긋난 전문용어를 바르게 고치는 표준화가 있다.

현재 국내에서는 국어기본법 제 17조에 따라 중앙행정기관에 전문용어 표준화협의회를 구성하고 각 기관에 '국어 책임관'을 지정하여 운영하도록 되어 있어 이를 기반으로 공공용어를 표준화하고 체계화해야 하지만 현행 전문용어 표준화 단순 지원으로는 한계가 발생하고 있다. 지정된 국어책임관들은 대부분 겸직하고 있고 이

마저도 인사이동 등으로 업무의 효율성과 연속성이 떨어진다. 이로 인해 전문 분야의 낯선 외국어가 일상 언어생활에 여과 없이 침투하여 정부 정책의 효율적 전달과 국민의 의사소통에 방해를 초래하고 있다. 또한 총괄적 국가적 관리 체계 미비로 인해 재난, 범죄 등 긴급 상황이 발생하였을 때 신속한 대응이 어려우며 세대에 따라 사용하는 용어 차이의 폭이 크게 벌어지고 있어 기술이나 지식을 공유하기 더욱 어려워진 실정이다. 산업계에서도 통일된 체계가 존재하지 않아 실제 사용할 때 동일한 사물에 대해 칭하는 명칭이 다른 경우가 많아 혼란을 야기하고 있다.

이를 보완하기 위해 국립국어원에서는 '전문용어 표준화 민관합동 총괄지원단'과 함께 쉬운 용어로 대체하는 개선 작업을 진행하고 있으며 부처별로 이루어지는 전문용어 표준화에 대한 검토 작업을 수행하고 있다. 지원 범위가 일부 기관, 일부 용어에 한정되어 있어 여전히 국민은 어렵고 낯선 전문용어로 인해 어려움을 겪고 있으며, 무분별한 외국어의 유입으로 더욱 가속화되고 있다.

따라서 행정부처에 전문용어 표준화를 맡겨두는 것이 아니라 국가적 차원에서 전문용어에 대한 표준화와 체계화를 확립할 수 있도록 제도적인 장치 마련이 필요하며 이를 총괄할 수 있는 '전문용어 통합 관리 기구'의 설치 또는 이에 준하는 시스템이 필요한 상황이다. 비대면 시대에 이를 총괄적으로 관리할 수 있는 온라인 협업 시스템과 각 기관별로 산재한 용어를 한 곳에서 검색, 활용, 연계 할 수 있는 보급 시스템 마련이 시급하다. 또한 시스템의 확대를 위해서는 언어 자원 기반의 신속하고 과학적인 발굴 및 용어 정비 방안 마련이 함께 구성되어야 한다.

II. 환경 분석

1. 환경 분석 개요

PEST 환경 분석 기법에 따라 정부 정책 환경(Political), 사회 환경(Social), 정보 기술 환경(Technological)을 분석한다[1]. 다양한 관점의 환경 분석으로 정보화 시사점을 도출한다.

[그림 1] 환경 분석 프레임워크

2. 정부 정책 환경

2.1. 국어기본법과 국어 발전 기본계획

국어기본법은 한국어 사용을 촉진하고 한글과 한국어의 발전·보전을 위해 제정되었으며 이전의 어문 규범, 국어 순화 등을 중심으로 한 정책에서 한발 나아가 한국어 보급과 국민의 국어 능력, 언어의 공공성 향상 등으로 그 초점을 옮기게 된다.

국어기본법 제17조에 국민이 쉽고 편리하게 사용할 수 있도록 국가가 각 분야의 전문용어를 표준화하고 체계화하여 보급하도록 함을 규정하고 있으며 전문용어의

1) PEST 분석에 경제 분석(Economical)이 포함되지만 본 사업의 특성과 규모를 고려하여 제외하였음.

표준화 및 체계화를 위하여 중앙행정기관에 전문용어 표준화협의회를 두어야 한다고 명시하고, 전문용어의 표준화 및 체계화 절차, 전문용어 표준화협의회 구성 및 운영에 필요한 사항은 대통령령으로 정하고 있다.

2007년부터는 국어 발전을 위한 종합 계획을 수립하여 5년마다 기본 계획을 세우고 중점과제, 추진 과제로 구성하여 국어 발전 기본계획을 시행하고 있다.

‘제1차 국어 발전 기본계획(2007~2011)’의 주요 과제는 국민의 국어 능력 향상을 위한 교육, 연수 체계 정비, 동북아지역 거점 기반 한국어 세계화 전략 추진, 다국어 지원 한국어 학습용 웹사전 편찬이며, 올바른 국어 사용을 위한 국가 언어정책의 확산, 남북 언어교류 확대 및 국제교류 협력망 구축, 소외계층을 위한 언어복지 시책 강화, 국어사용 환경 개선과 국민의 의사소통 증진, 국민의 국어 능력 증진 여건 조성, 언어 사용의 다양성 조사, 『표준국어대사전』의 정비 및 맞춤형 사전 편찬, 국어정보망 구축과 통합 정보시스템 운영, 국어 문화유산의 보전과 한글의 산업화, 국어문화 확산을 위한 홍보활동 강화를 10대 추진 과제로 삼아 수행하였다.

‘제2차 국어 발전 기본계획(2012~2016)’에서는 주요 5대 정책 과제로 품위 있는 언어생활을 위한 국민의 창조적 국어능력 향상, 공생·공영의 국어 문화 확산, 공공언어 개선으로 사회이익 증진, 한국어 보급으로 우리말 위상 강화, 우리말 문화유산 보전과 활용 기반 마련을 통한 국어 진흥 등이 포함되었다. 주로 정부나 지방자치 단체 등의 공공기관에서 사용하는 언어가 ‘쉽고 정확하여’ 사회 구성원 간의 원활한 소통을 방해하지 않고, ‘저속하거나 차별적이지 않아’ 사회 구성원의 삶의 질과 품격을 높일 수 있어야 한다는 것이 주요 내용이다[2].

‘제3차 국어 발전 기본계획(2017~2021)’의 주요 내용은 아래 [표 1], [표 2]와 같다.

2) 최혜원(2014). 공공용어 번역의 현황과 과제. 새국어생활 제24권 제2호

구분	주요 내용
목표	온 국민이 누리는 국어, 전 세계가 함께하는 한국어
추진 전략	• 언어 환경과 언어 현실을 반영한 국어정책 수립 • 한글의 가치와 한글문화의 국내외 확산 • 학습자 중심의 한국어교육 기반 강화 및 질적 도약 • 사회 통합을 위한 원활한 의사소통 환경 조성 • 국어 능력 향상과 바른 언어생활을 위한 여건 개선

[표 1] 제3차 국어 발전 기본계획 목표 및 추진 전략

추진 과제	실행 과제
수요자 중심의 언어정책 기반 조성	• 어문규범 현실화 • 국어사전의 개방적 운영 및 활성화 • 언어 정보 자원 구축 및 활용 • 국어 기본어휘 선정 및 어휘 등급화
바르고 편리한 언어 환경 지원	• 공공언어 개선 활성화 • 바른 언어문화 기반 조성 • 국민의 국어 능력 향상 지원 • 지역 언어문화 기반 국어문화원 활성화
국민 언어 통합을 위한 사회·문화적 환경 구축	• 남북 언어 통합 기반 구축 • 특수 언어 환경 개선 및 보급 확대 • 언어 취약 계층 지원 • 사회·지역 방언 정보의 구축과 활용
한국어 확산과 교육 기반 강화	• 국외 한국어 보급 대표 기관으로 세종학당 육성 • 한국어교육 체계화 및 기반 강화 • 한국어교원 자격 제도 운영 및 교원 연계망 구축 • 한국어교원 연수 과정 운영
한글문화 진흥 및 향유 확대	• 다양한 한글문화 자원의 수집 및 전시 • 한글문화 연구·교육 및 산업화 기반 구축 • 한글날 기념 및 한글문화 관련 포상 • 국립세계문자박물관 건립

[표 2] 추진 과제 및 실행 과제

2.2. 대구 정부통합전산센터 공공 클라우드 전환 설계

국가정보자원관리원에서 추진하고 있는 공공 클라우드는 스마트 전자정부 서비스의 일환으로 모든 행정기관의 정보 자원 수요를 모아 통합으로 관리하고 일괄 구축 및 공동 활용을 위해 필요한 만큼의 자원을 분배하고 서비스할 수 있도록 하는 기술 및 서비스이다.

국립국어원도 2022년까지 완공되는 대구 정부통합전산센터의 입주 대상으로 선정되어 있다. 소프트웨어는 오픈소스 기반으로 운영체제, 웹서버, 웹애플리케이션 서버, 데이터베이스관리시스템의 기준이 마련되어 있으며 하드웨어는 데이터 사용용량 및 사용자 접근 빈도수 등에 따라서 소형, 중형, 대형으로 구분하며 각 서버별 사용량에 적절한 업무 단위별로 시스템 분리를 수행해야 한다.

구 분		CPU(vCore)	메모리(GB)	비 고
소형	**소형(기본)**	**2**	**4**	**소규모 단순 업무**
	중소형	4	8	일반적인 내부 업무
중형	중형(기본)	6	12	중규모 Web, WAS, DB 서버
	중대형	8	16	중대규모 Web, WAS, DB 서버
대형	대형(기본)	12	24	대용량 CPU가 필요한 WAS, DB 서버
	초대형	16	32	대용량 CPU 및 고용량 메모리 DB 서버

[표 3] 대구센터 클라우드 하드웨어 사양

구 분	OS	Web	WAS	DBMS	비 고
기본	RHEL, Windows	-	-	-	OpenJDK (Windows JDK는 필요시 구성)
Web	RHEL	Apache	-	-	
WAS	RHEL	-	JBoss	-	
Web, WAS	RHEL	Apache	JBoss		
DB	RHEL	-	-	Cubrid, PostgreSQL, Altibase(x86), MariaDB	

[표 4] 대구센터 클라우드 소프트웨어 사양

따라서 향후 시스템 구축 시 공공 클라우드 전환 대상에 따르는 기준을 가지고 시스템이 인터넷망에서 서비스될 것인지 업무망에서 서비스될 것인지를 구분하고 Web, WAS, DB를 분리하여 3계층 구조 및 이중화 구조를 구성하며 업무 단위 시스템 분리를 통해서 향후 증설 및 확장이 쉽도록 시스템 아키텍처를 구성해야 한다.

3. 정보 기술 환경

3.1. 정보 기술 동향

□ IT 서비스의 변화

최신 기술에 대한 가트너[3] 하이프 사이클[4]은 모든 비즈니스에서 기술 혁명이 얼마나 빠르게 구매자, 공급자, 고객 관계를 재정의하는지를 보여 준다. 가트너는 올해 10개 전략기술로 인간을 중심으로 한 스마트 공간을 구현하는 것이 핵심이라고 했는데 이는 블록체인, 기계 학습, 범용 머신 인텔리전스, 스마트 워크 스페이스 등 많은 것을 포함한다.

2020년 하이프 사이클에 따르면 기업들이 비즈니스 모델을 재편하고 신흥 시장과 생태계에 접근할 수 있게 하는 다섯 가지 지배적인 경향으로, '증강 인간', '비

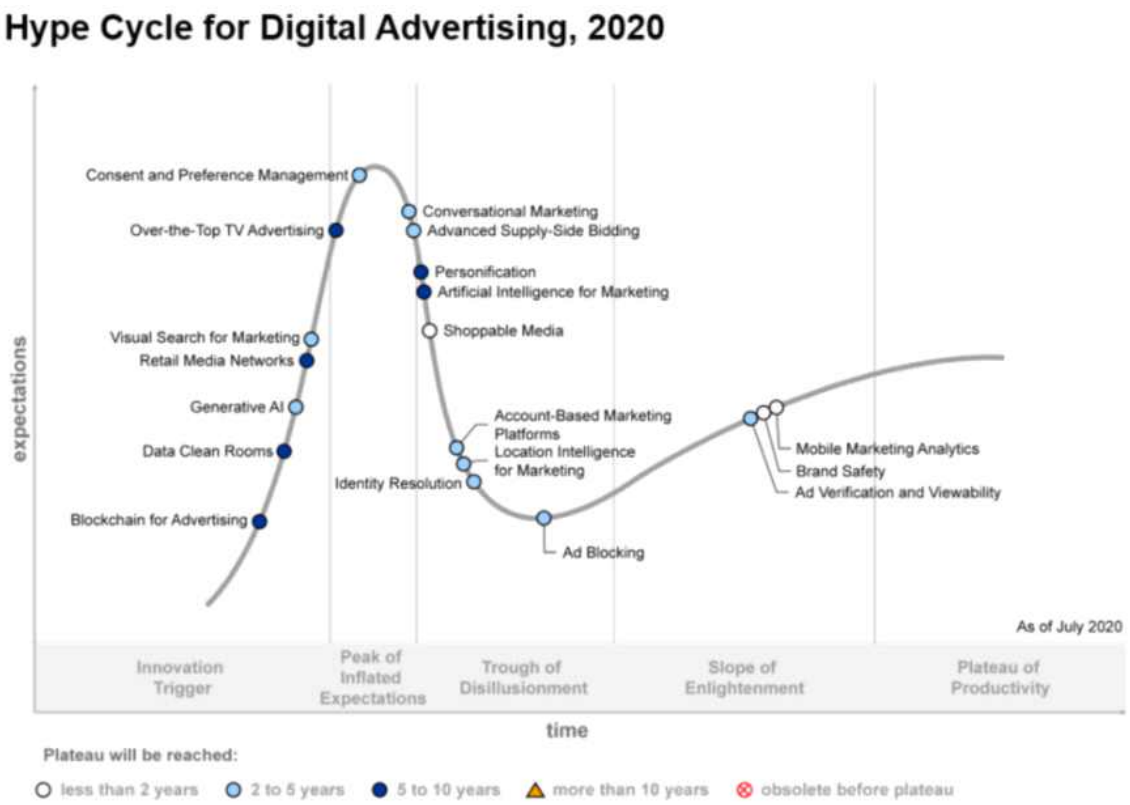
3) 미국 코네티컷주에 본사를 둔 IT 분야의 연구조사 기업이다. 2001년까지 가트너 그룹(TheGartnerGroup)으로 널리 알려졌으며 현재는 가트너로 불리고 있다. 다국적 IT 기업 및 각국의 정부기관 등을 주 고객으로 두고 있으며 설문 조사 부분의 높은 신뢰도로 공신력이 크다. 1979년에 설립되어 세계 75개국에 1,200여 명의 분석전문가와 고문을 포함 3,700여 명의 직원을 고용하고 있다.(출처: 네이버 지식백과)

4) 하이프 사이클(Hype Cycle)은 기술의 성숙도를 표현하기 위한 시각적 도구이다. 과대광고 주기라고도 한다. 미국의 정보 기술 연구 및 자문 회사인 가트너에서 개발하였다.(출처: 위키백과)
* 하이프 사이클은 5개의 단계로 이루어지며, 이는 기술의 성장 주기에 대응된다.

하이프 사이클

1. 기술 촉발(Technology Trigger): 잠재적 기술이 관심을 받기 시작하는 시기.
2. 부풀려진 기대의 정점(Peak of Inflated Expectations): 초기의 대중성이 일부의 성공적 사례와 다수의 실패 사례를 양산.
3. 환멸 단계(Trough of Disillusionment): 실험 및 구현이 결과물을 내놓는 데 실패함에 따라 관심이 시들해 짐.
4. 계몽 단계(Slope of Enlightenment): 기술의 수익 모델을 보여 주는 좋은 사례들이 늘어나고 더 잘 이해되기 시작. 2~3세대 제품들의 출시 시기.
5. 생산성 안정 단계(Plateau of Productivity): 기술이 시장의 주류로 자리 잡기 시작.

고전적 컴퓨팅 및 커뮤니케이션', '디지털 생태계', '고급AI 및 고급분석' 을 정의해 볼 수 있다. 2019년도에 AI가 가장 많이 언급되었는데 이에 더 세분된 기술들을 2020년도 발표한 하이프 사이클에서 언급하고 있다. 이는 2,000개 이상의 기술로부터 얻는 통찰력을 통해 반드시 주목해야 할 주요 신기술 및 동향을 간추려 낸 것으로, 기업들이 전략을 수립하는 데 상당한 영향력을 미칠 것이다.

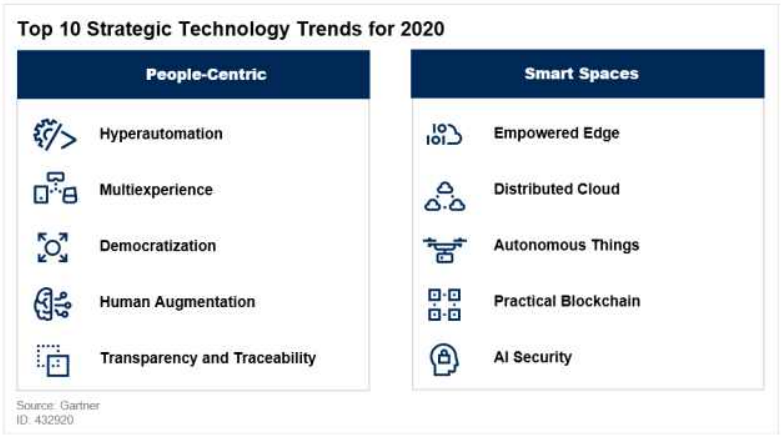


[그림 2] Hype Cycle for Emerging Technologies

*출처: Gartner, August 2020

□ 주목할 IT 동향

올해 가트너가 발표한 2020년도 전략 동향은 '초자동화', '다중 경험', '전문성의 민주화', '인간 증강', '투명성과 추적성', '자율권을 가진 에지', '분산형 클라우드', '자율 사물', '실용적 블록체인', '인공지능의 보안'으로 범주를 나눠 설명하고 있다.



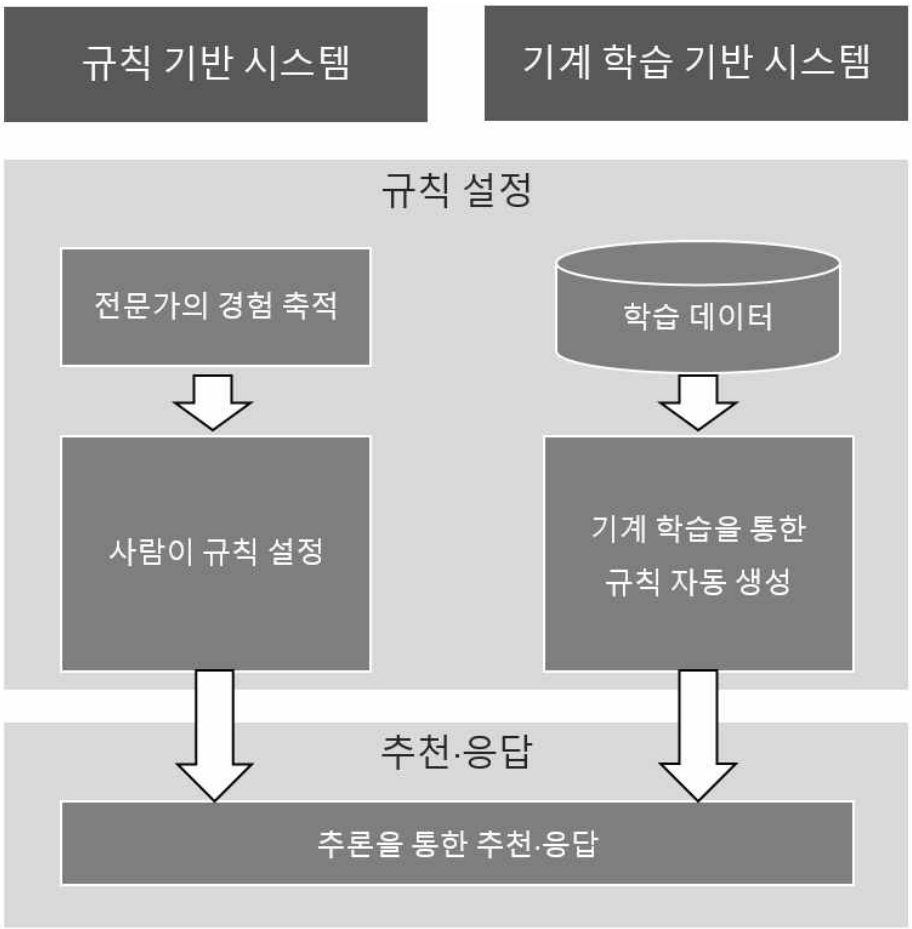
[그림 3] 2020 10대 전략 트렌드

3.2. 기계 학습 기술

인공 지능의 대표적인 방법이었던 전문가 시스템은 번역가, 의사, 변호사와 같은 전문가의 기법을 많은 수의 규칙으로 입력하여 문제를 해결하는 시스템이다. 하지만 그러한 규칙을 사람이 만들어서 등록해야 하는 어려움이 있고 사람조차 어떻게 규칙화할 수 있는지 모르는 문제가 있다. 대표적으로 음성 인식, 이미지 인식은 “열이 많이 나고 오한이 있고 구토 증상이 있으면 독감이다”라는 비교적 간단한 규칙으로 해결될 수 있는 문제가 아니다.[5)]

이런 문제점을 해결하고자 나온 방법이 기계 학습이다. 기계 학습은 많은 데이터를 컴퓨터에 제공하여 인간처럼 학습하게 함으로써 새로운 규칙을 얻는 방식이다.

5) 출처: 나무 위키



[그림 4] 규칙기반, 기계 학습 기반 시스템 비교

□ 기계 학습 방식

기계 학습은 학습 방식에 따라 3가지로 나눌 수 있다.

○ 지도 학습(Supervised Learning)

- 사람이 각각의 입력에 대해 정답을 달아 놓은 학습 데이터를 제공하면 컴퓨터가 학습하는 방식이다.
- 사람이 직접 정답 데이터를 제공하므로 정확도가 높지만 인건비 문제가 있고 구할 수 있는 데이터의 양도 적다는 문제가 있다.
- 대표적으로 자동 분류(classification), 음성 인식에 지도 학습 방식을 적용할 수 있다.

○ 비지도 학습(Unsupervised Learning)

- 사람이 만든 학습 데이터 없이 컴퓨터가 스스로 학습하는 방식이다.
- 정답 데이터 없이 문제를 푸는 방식이므로 학습이 맞게 됐는지 확인할 수 없는 문제점이 있지만 인터넷 시대에서 데이터의 양이 기하급수적으로 늘어남에 따라 활용하기 쉬워 선호되는 방식이다.
- 대표적으로 자동 군집화(clustering), 분포 추정에 비지도 학습 방식을 적용할 수 있다.

○ 강화 학습(Reinforcement Learning)

- 지도 학습, 비지도 학습은 인간이 컴퓨터에 학습 데이터를 제공했는가에 따른 것인데, 강화 학습은 조금 다른 형태로 현재의 상태에서 어떤 행동을 취하는 것이 최적인지 학습하는 방식이다.
- 행동을 취할 때마다 외부 환경에서 보상이 주어지는데 이러한 보상을 최대화하는 방향으로 학습이 진행된다. 그리고, 이러한 보상은 행동을 취한 즉시 주어지지 않을 수도 있는데, 이를 지연된 보상이라 한다.
- 이러한 이유로 강화 학습은 지도/비지도 학습과 비교하면 매우 어려운 방식이며 컴퓨터에 제대로 보상하는 문제와 신뢰 할당의 문제가 어려운 점이다.
- 대표적으로 바둑, 체스와 같은 인공 지능 게임에 강화 학습을 적용할 수 있다. 체스로 예를 들면 나와 상대의 말의 배치가 현재 상태고 '여기서 어떤 말을 어떻게 움직일까'가 행동이 된다. 상대의 말을 잡으면 보상이 주어지는데, 상대 말이 멀리 떨어져 잡을 때까지 이동 시간이 필요하므로 당장 보상이 주어지지 않는 경우가 있다.(지연된 보상)
- 따라서, 강화 학습에서는 당장 보상이 적더라도 나중에 보상이 합이 최대화하도록 행동을 취해야 하는데 상대방이 어떤 행동을 취할지 모르므로 미래를 고려하면서 가장 좋은 행동이 뭔지 여러 방식으로 고민해야 한다.

□ 기계 학습의 사용 사례[6]

○ 사기 방지

- 1억 5,000만 개의 디지털 지갑을 통해 연간 2,000억 달러 이상의 결제를 처리하는 '페이팔'(PayPal)은 온라인 결제업계의 선두 주자다.
- 이 정도 규모에서는 사기 비율이 낮다 해도 그 비용은 상당 규모에 이른다. 창업 초기에는 월별 사기 피해 금액이 1,000만 달러에 이르렀다.
- 페이팔은 이 문제를 해결하기 위해 최고의 연구원들로 팀을 꾸렸고 이 팀은 최신 기계 학습 기법을 사용해 사기성 결제를 실시간으로 식별하는 모델을 구축했다.

○ '타겟팅 디지털 디스플레이 광고'[7]

- 광고 기술 기업 디스틸러리(Dstillery)는 기계 학습을 사용해 버라이즌(Veriszon), 윌리엄스-소노마(Williams-Sonoma)와 같은 기업의 실시간 입찰 플랫폼에서 타겟팅 디지털 디스플레이 광고를 진행한다.
- 디스틸러리는 개인의 브라우징 내역, 방문, 클릭 및 구매에 대해 수집된 데이터를 사용해 한 번에 수백 개의 광고 캠페인을 처리하며 초당 수천 건의 예측을 실행한다.
- 덕분에 디스틸러리는 투자 대비 최적의 결과를 얻기 위한 타겟팅 광고에서 인간 판매담당자보다 훨씬 더 좋은 성과를 내고 있다.

○ 콘텐츠 추천

- 컴캐스트(Comcast)는 'X1'[8] 서비스 고객을 위해 각 고객의 이전 시청 습관을 기반으로 하여 실시간으로 개인 맞춤화한 콘텐츠를 추천한다.
- 컴캐스트가 운영하는 기계 학습은 수십억 개의 내용 기록을 사용해 고객별로 고유한 취향 프로필을 작성한 다음, 공통적인 취향을 가진 고객을 클러스터로 묶는다.

---

6) IT WORLD(2016년 1월). 머신러닝(기계 학습) 입문 가이드.
7) 사용자의 관심 분야, 방문 기록 따위의 정보를 이용하여, 인터넷 페이지를 이동할 때마다 그와 관련된 상품을 노출하는 광고를 말한다.(참조: 우리말샘 '리타겟팅 광고')
8) 'X1'은 컴캐스트(Comcast) 회사의 인터넷 기반 TV 서비스임

- 그 다음 각 고객 클러스터를 대상으로 가장 인기 있는 콘텐츠를 실시간으로 추적·표시해 고객이 현재 인기 있는 콘텐츠를 볼 수 있도록 한다. 기계학습을 이용하여 더 정확한 추천을 하게 되므로 고객 만족도가 높아지고 사용률이 증가한다.

○ 자동차 품질 개선

- 재규어 랜드로버(Jaguar Land Rover)의 신형 차량에는 60개의 '온보드 컴퓨터'(on-board computer)[9]가 탑재되며 이 컴퓨터는 2만 개 이상의 매트릭스를 기준으로 매일 1.5GB의 데이터를 생성한다.
- 재규어 랜드로버 엔지니어들은 기계 학습을 사용해 이 데이터에서 고객이 차량을 실제로 어떻게 다루는지를 파악해 낸다.
- 이렇게 얻은 정확한 사용 데이터를 통해 설계자는 부품 고장과 잠재적 안전 위험을 예측할 수 있다. 이는 예상되는 조건에 따라 적절히 차량을 제작하는 데 도움이 된다.

○ 유망 잠재 고객에 집중

- 판매담당자들은 최적의 판매와 마케팅 기회, 그리고 최적의 제품을 판단하기 위한 도구로 '구매 성향'(propensity to buy) 모델을 사용한다.
- 라우터부터 케이블 TV 상자에 이르기까지 방대한 제품을 보유한 시스코(Cisco)의 마케팅 분석팀은 몇 시간 만에 6만 개의 모델을 교육하고 1억 6,000만 명의 잠재 고객을 확보했다.
- 이 팀은 의사결정 트리(decision tree)부터 '그라디언트 부스팅'(gradient boosting)[10]까지 다양한 기법을 테스트함으로써 모델의 정확도를 대폭 개선했다. 이는 판매량 증가, 무익한 판매 전화 감소, 영업 담당자들의 만족도 향상으로 이어진다.

---

9) 차량, 냉장고, 전자시계 따위의 기계나 설비 안에 내장된 컴퓨터. 내부에 회로 기판이 장착되어 있으며, 기계나 설비를 제어하고 관리하는 데 쓰인다.(출처: 우리말샘)

10) 그라디언트 부스팅(gradient boosting)은 기계 학습 기술로 이전 학습의 결과에서 나온 오차를 다음 학습에 전달해 이전의 오차를 점진적으로 개선하는 기법이다.

○ 미디어 최적화

- NBC 유니버설(NBC Universal)은 국제 케이블 TV 배포를 위해 수백 테라바이트 용량의 미디어 파일을 저장한다. 또한 전 세계 고객을 대상으로 한 배포를 지원하기 위한 효율적인 온라인 자원 관리가 필요하다.
- NBC 유니버설은 기계 학습을 사용해 척도의 조합을 기반으로 각 항목의 미래 수요를 예측한다. 이런 예측을 기반으로 수요가 낮을 것으로 예상하는 미디어를 저렴한 오프라인 스토리지로 옮긴다.
- 기계 학습을 통한 예측은 파일 수명과 같은 하나의 척도에 기반을 둔 임의 규칙에 비해 훨씬 더 효과적이다.
- 결과적으로 NBC 유니버설은 고객 만족도를 그대로 유지하면서 전체 스토리지 비용을 줄이고 있다.

○ 의료 보건 서비스 개선

- 병원의 입장에서 환자의 재입실은 심각한 문제다. 환자의 건강과 복지도 문제지만 미국의 '의료 보험 공단'과 민간 보험사가 재입실 비율이 높은 병원에 불이익을 주기 때문이다.
- 따라서 향후 건강한 상태를 유지할 가능성이 충분히 높은 환자만 퇴원시키는 역량이 병원의 재무에 큰 영향을 미치게 된다.
- 캐롤리나 건강관리 시스템(Carolinas Healthcare System, CHS)은 기계 학습을 사용해서 환자의 위험 점수를 계산하고 병원 사례 관리자는 이를 바탕으로 퇴원 결정을 내린다.
- 이 시스템은 각 사례의 위험과 복잡성에 따라 환자에게 우선순위를 부여함으로써 간호사와 사례 관리자의 능률을 높여 준다. 그 결과 CHS는 재입실 비율을 21%에서 14%로 낮췄다.

□ 기계 학습의 성공 요소

기계 학습의 사용 사례에서 보듯이 기계 학습을 통해 인간의 사고와 분석적 한계를 뛰어넘어 엄청난 양의 이질적인 데이터로부터 가치를 발견할 수 있다. 하지만

기계 학습이 항상 좋은 결과를 내는 것은 아니다.

기계 학습으로 성공적인 결과를 내기 위해서 고려해야 할 핵심 요소가 있다[11].

○ 데이터가 많으면 더욱 정확해진다.

- 학습 데이터가 적을 경우 문제를 해결하기 위한 학습 모델의 복잡성을 뒷받침하지 못할 수 있다.
- 학습 데이터가 많으면 더욱 정확해지므로 표본이 아닌 우리가 가진 모든 데이터를 이용해야 한다.

○ 주어진 문제에 가장 적절한 기계 학습 방식을 선택하는 것이 핵심이다.

- 정확도가 높은 알고리즘인 'GBT'(Gradient Boosting Tree)는 업계 실무자들이 널리 활용하고 있는 인기 지도 학습 알고리즘이다. 하지만 그 높은 인기에도 불구하고 모든 문제를 위해 사용해서는 안 된다.
- 항상 그리고 가장 정확한 결과를 위해 데이터의 특성에 가장 적합한 알고리즘을 사용해야 한다.
- 여러 알고리즘을 비교 테스트하여 정확성을 검증할 필요가 있다.

○ 뛰어난 모델을 얻기 위해서는 알고리즘의 변수를 잘 선택해야 한다.

- 현대의 기계 학습 알고리즘은 변경할 수 있는 부분이 많다. 예를 들어, 인기 있는 'GBT 알고리즘' 단독으로도 트리(Tree) 크기 제어 방법, 학습률, 행이나 열의 샘플 채취 방법론, 손실 함수, 조직화 옵션 등을 포함해 최대 12개의 파라미터를 설정할 수 있다.
- 일반적으로 프로젝트에서는 각 파라미터에 대한 최적값을 찾아 주어진 데이터 세트에 대해 가장 높은 정확도를 얻어야 하는데, 그리 쉬운 일이 아

11) 머신러닝(기계 학습) 입문 가이드(IT WORLD)의 내용을 정리 요약함

니다.

- 직관과 경험이 도움이 되긴 하지만, 데이터 엔지니어는 최선의 결과를 위해 다수의 모델을 훈련하면서 교차 검증 점수를 파악하고, 다음에 시도할 파라미터를 결정하는 일을 고민해야 할 것이다.

○ 데이터의 특성을 잘 이해하고 이로부터 다양한 특징을 추출하는 방법을 고민해야 한다.

- 전문가와 데이터를 신중하게 검토하여 데이터의 생성 과정에 대한 통찰력을 얻는 것이 좋다. 이 과정으로 기록, 기능, 값, 표본 추출 등과 관련된 데이터 품질 문제를 식별할 수 있다.
- 수학적 변화 등 데이터 가공을 통해 데이터의 복잡한 특성을 잘 잡아내는 특징을 추출하여 학습에 활용하는 것을 고민해야 한다.
- 텍스트, 그래프 데이터, 이미지 등 종종 비구조화된 데이터를 활용해야 할 필요가 있다. 이러한 비구조화된 데이터로부터 다양한 특징을 추출하기 위한 방법을 파악해야 한다.

○ 기업 가치에 부합하는 적절한 함수의 선택이 중요하다.

- 거의 모든 기계 학습 알고리즘은 최적화가 중요하다. 기업의 특성에 기초해 최적화 함수를 적절히 설정하거나 조정하는 것이 기계 학습의 성공을 위한 핵심이다.
- 예를 들어, 'SVM'(Support Vector Machine)은 모든 오류 유형의 가중치가 동등하다고 가정함으로써 이진(binary) 분류 문제에 대한 일반화의 오류를 발생시킬 수 있다.
- 예를 들면 고장 감지 등 특정 유형의 오류 비용이 다른 것보다 더욱 중요할 수 있다. 이때, 가중치를 고려하기 위해 특정 유형의 오류에 더 많은 페널티를 더함으로써 'SVM' 손실 함수를 조정하는 것이 좋다.

○ 기업 문제를 기계 학습으로 해결하는 사례를 연구한다.

- 사기 감지, 제품 추천, 표적 광고 등 기업에서 중요하게 여기는 문제를 실제로 해결한 '표준' 기계 학습 방식이 있다.
- 이런 잘 알려진 문제뿐만 아니라, 덜 알려졌지만 예측 정확성이 더 높은 더욱 강력한 방식이 존재한다.
- 잘된 사례를 연구하여 참고할 필요가 있다.

4. 사회 환경 분석

전문용어 표준화는 국민이 쉽고 편리하게 소통할 수 있도록 공공 전문용어를 일정한 정비 기준에 따라 체계적으로 다듬어 알맞은 용어로 정하는 것이다[12].

초기에 전문용어 표준화는 학술적인 의미를 담는 전문용어를 대상으로 하였지만, 현재는 일반 언중에게 고시되거나 사용될 가능성이 많은 용어를 대상으로 하여 대중성을 가진 용어에 대해 표준화 작업을 수행하게 된다.

공공 정보 공개가 의무화되고 지식 공유 확산에 따라 '심장병', '인터넷', '콤플렉스'같이 일상용어인지 전문용어인지 구분하기 힘든 용어가 많아지고 전문가뿐만 아니라 일반인도 사용하는 전문용어가 점점 많아지고 있다.

12) 국립국어원 "2020 전문용어 표준화 안내서"

유형	예시
어려운 한자 전문용어	자동 제세동기(自動除細動器: 자동 심장 충격기)
	황천(荒天: 거친 날씨, 거친 바다)
낯선 외래 전문용어	펜데믹(pandemic: 감염병 세계적 유행)
	GIS(geographic information system: 지리 정보 시스템, 지리 정보 체계)
하나의 개념이 여러 용어로 사용되는 경우	탄탈럼 / 탄탈룸 / 탄탈륨 * 산업통상자원부 기술표준원에서 국가 표준으로 제시한 '분석화악용어(원소 이름)'에 따라 '탄탈럼'이 표준 용어임
	AED / 자동 제세동기 / 심장 세동 제거기 / 자동 심장 충격기
하나의 용어가 여러 개념으로 사용되는 경우	AI: 조류 인플루엔자(수의) / 인공지능(정보·통신) / 인공 수정(의학)
	PM: 맞춤 약물(약학) / 프로젝트 관리자(정보·통신) / 다발성 근염(의료) / 미세먼지(기상)
어법에 맞지 않는 전문용어	햇치(hatch) * 외래어 표기법에 따라 '해치'로 써야 함
	최대값 * 한글 맞춤법에 따라 '최댓값'으로 써야 함

[표 5] 표준화 대상 전문용어 예시(국립국어원 "2020 전문용어 표준화 안내서": 8쪽)

재난, 인터넷 신종 범죄 등 여러 사회 문제가 발생할 경우 대국민 정보 공유 등에 대응하기 위한 분야별 표준 용어와 용어 정의 등의 정보가 구축되어 있지 않은 경우가 많이 발생하고 있다. 특히, 국민 피해를 유발할 수 있는 용어인 스미싱이나 파밍 등의 용어들은 정보통신, 금융, 경찰행정 등에서 함께 쓰이는 용어로 해당 분야 간의 종합적인 검토가 필요한 용어임에도 불구하고 통합 연계 자료의 부재로 인해 즉각적인 대응이 어려운 실정이다.

최근 발생한 코로나19와 관련하여 지표환자, 초발환자 등과 같은 유사한 듯하지만, 전혀 다른 용어를 언론에서 병기하는 등의 오용 사례가 발생하지만 정작 관련 기관에서는 용어의 표준화된 정의가 부재하여 잘못된 용어를 바로 잡는데 많은 시간이 소요되고 있다.

한자 및 일본어 잔재에 익숙한 세대와 영어에 익숙한 세대 간의 소통 문제로 인해 기술 또는 지식 전수 및 신기술 재교육 등에 대해 불필요한 사회적인 비용이 많이 발생하고 있다.

[그림 5] 어려운 용어로 인한 세대 간 소통 문제(출처: 한국도로공사 보도 자료)

기관 또는 단체에서 구축하고 사용하는 전문용어의 경우 각자 만든 데이터베이스를 기반으로 하여 사용하고 있는데 이 전문용어도 필요 때문에 각자 만들어서 사용하므로 체계가 통일되어 있지 않다. 데이터베이스를 구성하는 항목이 다르고 분야 간 정보가 연계되어 있지 않아 여러 분야에 속한 공통된 용어의 경우 각 분야별로 사용하는 방법이 어떠한지에 대해 찾아보기가 어려운 경우가 많다.

남북한은 분단 이후 약 70여 년간 서로 다른 전문용어의 사용으로 인해 의학, 건설 등과 같은 전문 분야에서 소통하기 어려운 용어를 사용하는 경우가 많아지고 있다.

분야	남한 용어	북한 용어
의학	빈맥, 잦은맥박	속맥
	호산구, 산호성 백혈구, 호산성 백혈구	에오진 기호구
건설	공기 연행 콘크리트, EA 콘크리트	가스 콩크리트, 공기련행 콩크리트, 거품 콩크리트, 기포 콩크리트
	바닥판, 바닥 슬래브	층막, 층막판

[표 6] 남한과 북한 전문용어 비교 예시

이를 해결하기 위해서는 여러 분야에 걸쳐 교류와 협력을 확대하고 전문용어의 표준화가 선결되어야 향후 남북통일 시에도 표준화된 용어를 통해 의사소통이 원활하게 이루어질 수 있을 것이다.

5. 환경 분석 종합

정부 정책 환경, 정보기술 환경, 사회 환경 분석을 통해 시사점을 도출하였다.

○ 정부 정책 환경

- 국어기본법<제17조>에 따라 국민이 전문용어(공공용어)를 쉽고 편리하게 사용하도록 표준화·체계화하여 보급해야 하는 의무가 있음
- 국어 발전 기본계획의 하나인 사회 통합을 위한 원활한 의사소통 환경 조성하기 위하여 전문용어를 표준화하고 체계적으로 관리할 수 있는 총괄적 체계 구축이 필요함
- 2022년도 대구 정부통합전산센터 입주를 예정하고 있으므로 공공 클라우드 아키텍처를 준수하여 시스템을 구축해야 함

○ 정보기술 환경

- 기계 학습과 심층 학습을 중심으로 한 지능화를 강조한 인공지능(AI)이 부각되고 있음
- 고급 인공지능(AI)과 함께 보다 복잡하고 정교한 기술과 도구를 사용하여 데이터를 자동으로 분석하는 등 지능정보기술이 지속적으로 영향력을 발휘할 것으로 보임
- 규칙 기반의 알고리즘을 넘어서 이해 학습 예측 및 적응을 하며 스스로 가동되는 자율 시스템을 만들며 기계 학습, 에지 인공지능(Edge AI), 인공지능 플랫폼 등 성공 핵심 요소를 고려해야 함

○ 사회 환경

- 공공 정보 공개의 일상화 등을 통해 전문용어가 다양해지고 전문가뿐만 아니라 일반인들도 전문용어를 접할 기회가 많아지고 있음
- 공공용어 관리 체계 부재로 인해 수시로 변하는 사회 문제에 대해서 즉각적인 대응이 어려움
- 각 분야 또는 기관별로 자체적인 데이터베이스를 보유하고 있어 상호 참조가 불가능하며, 데이터베이스별로 구성 항목이 상이하여 일관성이 없으며, 용어를 이해하고 사용하기 위한 기본 정보인 용어 해설, 사용 예시 등의 정보를 확인하기가 어려움
- 남북통일을 대비하여 전문용어에 대해 교류 및 협력이 필요함
- 전문용어의 표준화 작업을 수행하여 사용자의 혼란을 줄이고 분야 내, 분야 간 지식 교류를 활성화해야 함

III. 현황 분석

1. 현황 분석 개요

국내외 공공용어의 관리 및 추진 현황 분석을 통하여 이슈 사항을 정리하고 국가적 차원에서 공공용어 관리 및 향후 발전 시사점을 도출한다. 국내외 현황 분석은 프랑스, 캐나다, 중국의 공공용어 관리 및 추진 현황을 분석하고 국내의 공공용어 및 전문용어, DB 구축 현황을 조사하고 분석한다.

2. 해외 공공용어 관리 현황

2.1. 개요

공공용어를 국가 차원에서 관리하고 운영하는 대표적인 사례인 캐나다, 프랑스, 중국의 사례를 분석하고 각 국가별 추진 과정, 관련 법령, 연구기관 현황, 연구 분야와 전문용어 관리 현황을 살펴본다.

2.2. 캐나다 전문용어 정비

□ 추진 과정

○ 1961년 퀘벡 의회법으로 ‘프랑스어청(Office de la langue française)’을 설립하고 퀘벡에서 프랑스어의 질을 보장하는 역할 위임

○ 1969년 ‘공식어 진흥법’인 ‘63호 법’, 1974년 ‘공식어법(22호법)’, 이후 현재까지 시행되고 있는 1977년 ‘프랑스어 헌장(101호법)’을 통해서 프랑스어청은 정치적 기구로서 프랑스어를 직업어(langue de travail)로 정착시키는 조치를 취하고 적극적인 언어정책과 전문용어 정비 작업

□ 관련 법령

○ 1969년 ‘공식어 진흥법’인 ‘63호 법’, 1974년 ‘공식어법(22호법)’, 이후 현재까지 시행되고 있는 1977년 ‘프랑스어 헌장(101호법)’

○ ‘프랑스어 헌장(101호법)’은 가장 강력하고 배타적인 언어법으로서 모든 공공부문과 공적인 역할을 갖는 모든 민간 부문에서 표준화된 프랑스어 및 정비된 용어를 쓰도록 하는 법안

□ 연구 기관

○ 프랑스어청과 각 행정부처 내부에 ‘전문용어 위원회’를 두고 부처 관련자 및 전문가들 스스로 용어를 표준화

○ 모든 산업 부문의 기업들과 각 정부부처가 분야별로 전문용어를 정비하고 프랑스어 관련 연구와 조사를 수행하며 정부 대상으로 자문하며 프랑스어화 증서 프로그램을 관리하는 역할을 수행

□ 연구 분야

○ 공공기관, 경제생활(직업 부문), 교육 부문

- 사법언어
- 행정언어
- 공공 목적의 기업 및 직업의 언어
- 상업/경영 분야, 소비재와 관련된 언어
- 교육 언어

□ 전문용어 현황

○ 전문용어 대사전(GDT) 서비스: 300만 개 이상의 전문용어 서비스

- 전문 분야 수: 167개 영역
- 수록 용어 수: 약 300만 개

[그림 6] GDT에서 제공하는 직업분야별 용어집 예시

○ TERMIUM Plus: 캐나다 번역청(연방정부 산하)이 서비스하고 있는 세계에서 가장 방대한 전문용어 데이터베이스

- 전문 분야 수: 24개 대분류, 각 대분류에 10~12개 중분류, 약 250개의 분야
- 수록 용어 수: 약 400만 개

2.3. 프랑스 전문용어 정비

□ 추진 과정

○ 1952년 '과학언어평의회(Conseil du Langage scientifique)'를 결성하고 1955년부터 과학한림원 내부에서 과학언어자문위라는 이름으로 활동

○ 1952년 일반 대중에게 언어 관련 정보를 제공하는 잡지 『생활과 언어(Vie et langage)』 창간

○ 1954년 '프랑스어기술용어연구위원회(Comité d'étude des termes techniques français)' 결성

○ 1957년 ‘프랑스어어휘협회(Office du vocabulaire français)’ 창설

○ 1958년 민간단체 ‘프랑스어 수호’ 결성

○ 1964년 5개국(프랑스, 벨기에, 캐나다, 스위스, 모리셔스)의 13개 단체가 ‘세계프랑스어연맹(Fédération du français universel)’ 결성

○ 1966년 대통령령으로 ‘프랑스어 수호와 확산을 위한 고위위원회’가 설치되었으며 이후 1972년 중앙행정부처 소속 전문용어위원회들을 각 부처에 설치하고 장관명령에 따라 여러 부처에서 8개의 전문용어위원회를 신설하고 1973년 1월 표준 용어들을 처음 공보에 게시

○ 1986년 총리령에 따라 전문용어위원회의 작업을 조율하는 ‘전체위원회(Commission générale)’를 설치하고 각 행정부처의 활동에 대한 ‘전문용어 연차보고서’ 작성, 총리에게 제출

○ 1996년 총리령에 따라 전문용어·신어 총괄위원회와 전문위원회들을 설치하고 각 부처 장관은 전문용어·신어 책임관을 임명하여 전문위들은 표준 용어를 총괄위원회에 제출, 총괄위원회는 심사·선정한 후 프랑스한림원의 동의를 얻어서 해당 부처 장관에게 통보하며 전문위가 선정하고 프랑스한림원이 동의한 용어 목록은 ‘공보’에 고시

□ 관련 법령

○ ‘프랑스어 풍부화 법’(Décret n. 2015-341 du 25 mars 2015 modifiant le décret n.96-602 du 3 juillet 1996 relatif à l'enrichissement de la langue française)

□ 연구 기관

○ ‘프랑스어 풍부화위원회’는 프랑스 국무총리 산하 직속기구로서 프랑스 어휘의 보강과 풍부화를 위한 전문용어 제정 및 정비 기구

○ ‘프랑스어와 프랑스의 언어들 총대표부(DGLFLF)’(우리나라의 ‘국립국어원’에 해당), 13개의 정부부처에 분배되어 있는 19개의 과학기술 분야의 전문위원회의 상시 합동 운영 체계, 프랑스 한림원(아카데미 프랑세즈)과 대학연구소들, ‘프랑스표준화기구(AFNOR)’, 프랑어권 국가들의 언어정책 전문가들과의 협력으로 작업

○ 프랑스 한림원의 최종 감수 후 ‘프랑스 공보(Journal officiel)’에 공포되고 이후 모든 행정기구, 정부부처에서 사용이 의무화됨.

□ 연구 분야

○ 항공분야, 화학분야, 생물학분야, 국방분야, 재무분야, 미디어분야, 자동차분야, 법분야, 식품분야, 농업분야, 우편분야, 음향분야 등 총 76개 분야 대상

□ 전문용어 현황

○ 용어 서비스: 모든 용어는 문화부의 ‘프랑스 용어(france terme)’ 사이트에서 검색

- 전문 분야 수: 76개 분야
- 수록 용어 수: 약 8,000개

[그림 7] ‘FranceTerme’ 누리집의 첫 화면

2.4. 중국 전문용어 정비

□ 추진 과정

○ 1915년: 장쑤교육회의 물리화학교수연구회에서 최초로 물리학과 화학 전문용어를 심의결정. 점차 수학, 동물학, 식물학, 의학까지 분야를 확대하여 전문용어의 심의결정을 확대

○ 1923년: 전문용어집 '광물암석 및 지질명사 집요(矿物岩石及地质名词辑要)'를 펴냄

○ 1931년까지: 14개 분야의 전문용어를 심의결정

○ 1987~2019년: 전문용어 통일 기구인 '전국 자연과학 명사 심의결정위원회'에서 천문학, 토양학, 대기과학, 언어학, 생리학, 의학, 자동화 등 126개의 분야에 대해 전문용어를 심의 결정해 발표

○ 1989~1992년: 중국 본토 외의 중화권 지역을 위한 해외판도 번체자[13]로 심의 결정해 발표

□ 연구 기관 및 연구 실적

○ '전국 과학기술 명사 심의결정위원회'에서 전문용어 표준화를 총괄

○ 1909년: 최초의 과학기술 전문용어 심의 결정 및 통일 기구인 과학명사 편정관(科学名词编定馆) 설립(청 정부 시기)

○ 1918년: 청 정부의 과학 부처인 중국과학사에 의해 과학명사 심의결정위원회가 성립

○ 현 정부가 들어선 후로, 중국과학원 편역국(编译局)에서 청 정부의 국립편역관이 제정한 분야별 전문용어 초안을 인수인계받음

13) 중국에서 전통적으로 써 오던 방식 그대로의 한자 글씨체. 현대 중국의 간체자에 상대하여 이르는 말이다. (출처: 우리말샘)

○ 1950년: 학술명사 통일사업위원회 설립. 자연과학, 사회과학, 의약위생, 시사문학, 예술 등 다섯 개의 분과를 둠

- 각 분과는 다시 하위 소분과를 두었는데, 자연과학 분과의 경우 천문학, 수학, 물리학, 화학, 동물학, 식물학, 지질학, 지리학, 지구물리학, 엔지니어, 농학 등 소분과로 나뉨
- 전문용어의 심의, 결정은 분과별로 수행. 자연과학 분과의 예를 들면, 중국과학원은 여러 관련 학회와 연구기관으로부터 제보받은 전문용어 후보를 선별하며, 이에 대해 저명한 과학자 150인으로 구성된 문화교육위원회에서 최종 심사하고 결정

○ 1956년 문화교육위원회가 해산. 중국과학원은 편역출판위원회의 산하에 명사실(名词室)을 두어 전문용어의 통일과 심의결정을 주관함

○ 1960년대 초: 중국과학원 편역출판위원회 명사실이 자연과학 명사 편정실로 변경

○ 1960년대 중반~1970년대 중반: 중국의 '문화대혁명' 시기, 정치 환경으로 말미암아 전문용어 정비 사업은 10여 년간 중단

○ 1978년: 중국과학원의 주관하에 전국자연과학명사심의위원회가 설립 및 전문용어 표준화 사업 재개

○ 1985년: 전문용어 통일 기구인 '전국 자연과학 명사 심의결정위원회'가 성립

○ 2020년 현재: '전국 과학기술 명사 심의결정위원회'로 확대되어 있으며 95개의 분야별 전문용어 심의결정위원회를 두고 있고 수천 명의 학자들이 심의결정 사업에 참여하고 있음

□ 연구 분야

○ 천문학, 물리학, 생물화학, 전자학, 농학, 의학, 언어학, 교육학 등 95개 분야별 전문용어 심의결정위원회를 두고 심의결정 사업에 참여 중

□ 전문용어 현황

○ TermOnline 서비스: 온라인 전문용어 데이터베이스(termOnline) 검색 서비스 운영

- 전문 분야 수: 95개 분야
- 수록 용어 수: 약 50만여 개

[그림 8] ‘termonline’ 첫 화면 갈무리

3. 국내 공공용어 관리 현황

국내 공공용어의 관리 현황을 살펴보고 해외 사례와 비교하여 현재의 문제점을 개선하고 향후 국가적 차원에서 공공용어의 발전 방향을 도출한다.

□ 추진 과정

○ 1960년대 이전: 해방 후의 혼란기와 6.25 전란으로 정부와 학계의 무관심

○ 1976년: 과학기술용어집 제1집 발간, 과학기술단체총연합회 (과학기술처 보조)

○ 1978년: 의학용어집 제2집 발간, 과학기술단체총연합회

○ 1998년: 과학기술용어집 발간(약 22만 개), 한국과학기술한림원

○ 2000년: 개발환경 구축 및 기본자료 집성, 전문용어언어공학연구센터(한국과학기술원)

○ 2003년: 전문용어 집성(과학기술분야용어), 전문용어언어공학연구센터(한국과학기술원)

○ 2005년: 핵심과학기술용어집 발간(약 2만 개), 한국과학기술 한림원(과학기술용어의 통일화 작업을 그 창립의 과제로서 채택함)

○ 2005년: 영·한 기초과학표준용어집 발간(약 3만 개), 전문용어언어공학연구센터(한국과학기술원)

○ 2006년: 전문용어 기초용어 데이터베이스 구축(전산학용어, 전자전기공학용어, 기계공학 분야 용어), 전문용어언어공학연구센터(한국과학기술원)

○ 2008년: 국방과학기술용어사전 발간(7,500개), 국방기술품질원

○ 2011년: 국방과학기술용어사전 발간(15,000개), 국방기술품질원

○ 2013년: 방위사업용어사전 발간(5,000개), 방위사업청

□ 관련 법령

○ 국어기본법 제17조(전문용어의 표준화 등) 3항 신설

- 시행령 제12조(표준화협의회의 구성 및 운영)가 2019년 1월 1일부터 시행
- 전문용어 표준화협의회를 구성하고 표준화 및 체계화 절차를 법적인 절차에 따라 수행

□ 연구 기관

○ 1960년대: 일부 학회가 자기 학문 분야의 용어에 대한 검토 시작

○ 이후 1970년대 과학기술단체총연합회에서 지속적으로 과학기술 용어집을 발간

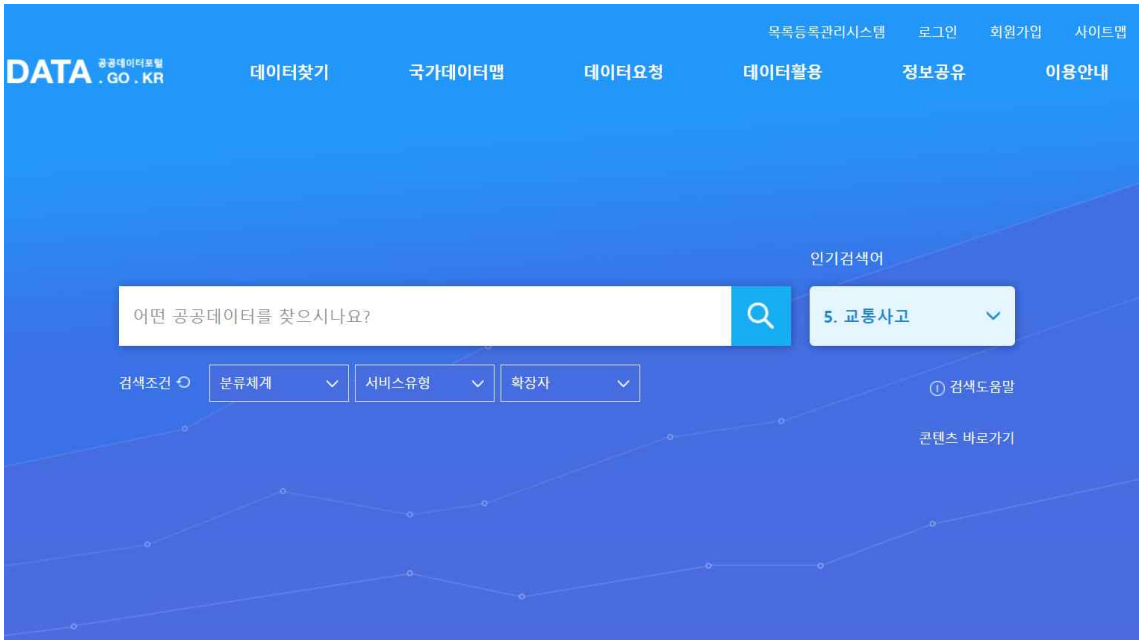
○ 2000년대 전문용어언어공학연구센터에서 과학기술 전문용어 데이터베이스 구축 및 용어집을 발간하였으나 2006년 이후 예산지원 종료로 인해 현재 국가 또는 민간에서 수행하는 연구기관 전무

□ 전문용어 현황

○ 공공데이터 포털: 공공데이터 포털은 공공기관이 생성 또는 취득하여 관리하

는 공공데이터를 한곳에서 제공하는 통합 창구로 국민이 쉽고 편리하게 공공데이터를 이용할 수 있도록 파일데이터, 오픈 API, 시각화 등 다양한 방식으로 자료를 제공

- 용어 관련 데이터 수: 약 91만 건



[그림 9] 공공데이터 포털 첫 화면 갈무리

□ 기타 현황

○ 보건의료정보표준관리시스템: 시스템 간 상호운용성 보장을 위하여 다양하게 표현되는 용어들에 대해 같은 의미를 지원할 수 있도록 개념화(대표어·동의어)한 용어체계로 사회보장정보원에서 관리 중인 서비스

- 데이터베이스 현황: 통합 용어 테이블을 구성하여 개념코드(216,698개)와 용어코드(321,217개)로 구분하고 용어의 한글 표기, 영문명 등을 관리
- 용어의 제안, 투표, 검증 등 용어 선정을 위한 표준활동 수행
- 회원이 의료용어 매핑 도구를 사용하여 표준화 활동에 참여함

[그림 10] 보건의료정보표준 첫 화면 갈무리

○ 전문용어 활용앱: 공공데이터 포털 > 데이터 활용 > 공공데이터 활용사례

- 정부부처에서 제공하는 전문용어를 기반으로 하여 사용자가 쉽게 접근할 수 있도록 모바일 앱으로 제공

[그림 11] 공공데이터 포털 누리집 내 공공데이터 활용 사례를 '용어'로 검색 결과

4. 국내외 공공용어 관리 현황 분석 종합

해외 공공용어 관리 사례 분석과 국내 공공용어 관리 현황 분석을 통해 현재의 문제점을 도출하고 국가적 공공용어 관리의 필수 요건을 도출하였다.

○ 해외 공공용어 관리 사례 분석

- 캐나다 사례 분석
  - 법령으로 용어를 관리하는 '프랑스어청'을 설립하고 지속적으로 법안을 개정하여 언어정책과 전문용어를 정비함
  - 공공 부문뿐 아니라 모든 민간 부문에서도 강제성을 가지고 산업 부문의 기업에서도 사용하도록 프랑스어화 증서 프로그램을 관리함
  - 일정 규모 이상의 기업에서는 내부의 프랑스어 위원회를 설치하도록 법령으로 강제하고 있으며 이를 위해 국가에서 지원금 제공
  - 지속적인 프랑스어청의 운영으로 현재 167개 영역의 300만 개 이상의 전문용어를 발굴하고 표준화함

- 프랑스 사례 분석
  - 1952년부터 과학한림원 내부에 과학언어자문위라는 이름으로 기구를 결성하여 대중에게 언어 관련 정보 제공 잡지를 발간하며 지속적으로 민간단체 및 국제단체 등 활동을 유지함
  - 1966년부터 대통령령으로 '고위위원회'를 설치하여 국가적 차원에서 관리가 이루어짐
  - 현재 '프랑스어 풍부화법'을 통해 프랑스 공보를 통해 공포된 용어들은 모든 행정기구, 정부부처에서 강제 사용됨
  - 모든 용어를 문화부에서 운영하는 프랑스 용어(FranceTerme) 사이트를 통해 서비스함
  - 현재 정비된 용어는 76개 분야 8천여 개가 공개되어 있음

- 중국 사례 분석
  - 1915년부터 장쑤교육회의 물리화학교수연구회에서 최초로 물리학과 화학 전문용어를 심의결정. 점차 수학, 동물학, 식물학, 의학까지 분야

를 확대하여 전문용어의 심의결정을 확대하였으며 이후 지속적인 국가 차원의 전문용어 관리체계에 돌입하여 간체자와 번체자까지 심의로 결정해 발표

- ‘전국 과학기술 명사 심의결정위원회’에서 전문용어에 대한 전반적인 표준화를 총괄하고 있고 95개 분야별 전문용어 심의결정위원회를 두고 있으며 각 분야별 상세 분야로 120개의 분야별 전문용어를 매년 발표하고 있음
- 온라인 전문용어 데이터베이스(termOnline) 검색 서비스를 운영하고 있으며 100여 개 분야 전문용어를 총망라함
- 일반 사용자로부터도 광범위하게 전문용어 표제어와 관련된 정보를 제보받는 창구로도 활용되고 있음

○ 국내 공공용어 관리 사례 분석

- 국어기본법 제17조(전문용어의 표준화 등) 3항 신설
  - 시행령 제12조(표준화협의회의 구성 및 운영)가 2019년 1월 1일부터 시행
  - 전문용어 표준화협의회를 구성하고 표준화 및 체계화 절차를 법적인 절차에 따라 수행해야 하며 중앙부처에는 각 부처별로 국어책임관을 두고 수행해야 하나 겸직 및 인사이동 등으로 인해 업무의 연속성을 가져갈 수 없고 전문용어 표준화 업무에만 투입될 수 없어서 업무를 수행하는 데 있어 한계가 있음

- 공공용어 표준화 현황
  - 공공데이터 포털을 통해 일부 부처 및 산하기관에서 보유한 데이터를 공개하고 있으나 항목 및 내용 등이 공개한 기관 중심으로 데이터 가 구성되어 있어 사용자에게 필요한 정보가 누락되어 있는 경우가 많고, 데이터의 제공 방식도 표준화되어 있지 않아 기관별로 다 다른 것을 확인할 수 있음
  - 필요에 따라 기관에서 자체적으로 전문용어를 제공하는 경우도 있으나 지속적으로 유지되지 못하거나 관리되지 않아 실질적으로 쓰이지 못하

는 경우도 다수 존재함

5. 공공기관 용어 자료 현황

정부 및 산하기관에서 보유하고 있는 언어 자원 중 공공데이터 포털 및 기관별 자체 누리집 등에 공개된 언어 자원들을 살펴보면 다음과 같다. 여기에 등장하는 기관은 51곳이고 자료집은 73개이며, 용어는 총 918,184개이다.(2020년 12월 기준)

행정부처명		자료집명	자료 형태	건수	항목 정보	이용허락범위
중앙행정부처	기획재정부	시사경제 용어사전	도서 모바일앱 기관 누리집	2,982건	주제, 용어, 설명	출처 표시 + 변경금지
	통일부	북한용어 사전	기관 누리집	3,408건	표제어, 설명	출처 표시 + 변경금지
	농림축산식품부	농업용어 사전	기관 누리집	104,392건	표제어, 설명, 다국어	저작자 표시
		축산용어 사전	기관 누리집	1,499건	표제어, 설명	저작자 표시
		용어집	기관 누리집	110건	분류, 기존용어, 한자어(원어), 쉬운 용어	저작자 표시
	산림청	산림임업 용어사전	기관 누리집	11,179건	국문명, 영문명, 한자, 용어설명	출처 표시 + 변경금지
	산업통상자원부	지식경제 용어사전	모바일앱	2,791건	표제어, 설명	출처 표시 + 변경금지
		통상 용어 정보	공공 데이터 포털	685건	표제어, 영문명, 설명	저작자 표시
		금속표준 용어	공공 데이터 포털	14,349건	표제어, 영문명	이용허락범위 제한 없음
	보건복지부	통계용어 사전	기관 누리집	407건	표제어, 영문명, 설명	이용허락범위 제한 없음
	환경부	환경용어 사전	기관 누리집	1,500건	표제어, 영문명, 설명	저작자 표시
	행정안전부	사회재난 핵심용어 집	기관 누리집	2,012건	표제어, 영문명, 설명	저작자 표시

행정부처명		자료집명	자료 형태	건수	항목 정보	이용허락범위
	국토교통부	국토해양 용어사전	기관 누리집	1,068건	표제어, 설명	저작자 표시
		용어사전 (철도산업 정보센터)	기관 누리집	10,631건	표제어, 외국어표기, 설명, 관련용어	확인 불가
		공간정보 용어사전 (국토지리 정보원)	기관 누리집	2,073 건	용어, 한자, 영문, 정의	저작자 표시
		항공정보 용어사전	공공 데이터 포털	8,449건	용어, 한글명, 약어, 해설	이용허락범위 제한 없음
		지적용어 사전 (지적측량 바로처리 센터)	기관 누리집	1,867건	표제어, 한자명, 영문명, 설명	출처 표시 + 변경금지
		해양용어 사전 (해양GIS 포탈)	기관 누리집	1,155건	표제어, 영문명, 설명	저작자 표시
		토지이용 용어사전	기관 누리집	630건	용어, 용어해설	출처표시 + 변경금지
		항공사고 관련 용어	공공 데이터 포털	73건	용어, 설명	이용허락범위 제한 없음
	국방부	국방데이터 표준단어	공공 데이터 포털	14,619건	단어명, 물리명, 영문명, 금칙어	이용허락범위 제한 없음
		국방과학 기술 용어정보	공공 데이터 포털	16,020건	구분, 용어(한글), 약어, 용어(영어), 출처, 설명	이용허락범위 제한 없음
	국세청	국세통계 용어사전	기관 누리집	1,633건	한글명, 한자명, 영어명, 해설	확인 불가
	방위사업청	방위사업 용어사전	공공 데이터 포털	5,032건	표제어, 대분류,	저작자 표시

행정부처명		자료집명	자료 형태	건수	항목 정보	이용허락범위
					영문, 설명, 출처, 출처1, 중분류	
	관세청	관세용어 사전	기관 누리집	4,889건	표제어, 설명	저작자 표시
	통계청	국세통계 용어사전	기관 누리집	1,327건	표제어, 해설	확인 불가
	법제처	법령 정의사전	기관 누리집	69,796 건	용어, 정의, 출처	이용허락범위 제한 없음
	기상청	대기과학 용어사전	도서	21,000여 건	용어, 해설	확인 불가
		항공기상청 용어사전	기관 누리집	173건	용어, 해설	출처 표시 + 변경금지
기타	금융감독원	금융용어 사전	기관 누리집	538건	용어, 영문명, 해설	출처 표시 + 변경금지
	한국지역난방공사	지역난방 용어 사전	공공 데이터 포털	1,038건	용어, 분야, 영문명, 설명	이용허락범위 제한 없음
	한국전력공사	전력관련 용어 정보	공공 데이터 포털	4,373건	용어, 용어설명	이용허락범위 제한 없음
		전력용어 순화 결과	공공 데이터 포털	327건	현행용어, 순화어, 용어의미 및 사용예시	이용허락범위 제한 없음
		용어 사전	기관 누리집	3,873건	용어, 용어설명	출처 표시 + 변경금지
	한국산업기술평가관리원	지역발전 정책용어 사전	공공 데이터 포털	21건	용어, 생성배경, 용어 설명, 현재 용어의 사용, 해외 사례, 그림, 참고자료 문헌	저작자 표시
		산업기술 혁신사업 관련 용어사전	공공 데이터 포털	77건	용어, 해설	이용허락범위 제한 없음
	한국환경산업기술원	환경 및 무역 관련 용어정의	공공 데이터 포털	30,847건	용어, 해설	이용허락범위 제한 없음
	한국조폐	조폐기술	공공 데이터	343건	용어,	이용허락범위

행정부처명	자료집명	자료 형태	건수	항목 정보	이용허락범위
공사	용어 정보	포털		영문명, 해설	제한 없음
사회보장정보원	보건의료 정보용어 표준	기관 누리집	321,217건	용어코드, 개념코드, 영문명, 한글명, 버전	저작자 표시
한국법제연구원	법령용어 정보	기관 누리집	1,184건	한글용어,한자용어, 용어설명	저작자 표시
한국정보통신기술협회	정보통신 용어사전	기관 누리집	24,590건	표제어, 한자명, 영문명, 출처, 동의어, 설명	출처 표시 + 변경금지
한국과학기술정보연구원	과학기술 데이터 용어 정보	공공 데이터 포털	61,954건	분류1, 아이디,용어1, 용어2,값, 서브값,용어1_언어, 용어2_언어 분류2, 아이디,용어1, 출처,용어2, 값,서브값, 용어1_언어, 용어2_언어	이용허락범위 제한 없음
국가보훈처	보훈행정 용어집	공공 데이터 포털	164건	대분류, 표제어, 뜻, 참고	저작자 표시
한국생산기술연구원	엔지니어링플랜트 표준용어집	공공 데이터 포털	1,107건	국문(원어), 영문명	저작자 표시
국사편찬위원회	한국역사 용어시소러스 정보	공공 데이터 포털	64,212건	최상위용어, 용어명, 관계, 용어명(한자	이용허락범위 제한 없음

행정부처명		자료집명	자료 형태	건수	항목 정보	이용허락범위
					), 용어 설명, 용어 사용년도, 용어 사용시기, 용어 설명	
	한국수력원자력	표준중수로용어집	공공 데이터 포털	6,531건	영문용어, 한글용어	이용허락범위 제한 없음
		원자력발전 남북한 용어사전	공공 데이터 포털	23,519건	영문용어, 한글용어, 북한용어	이용허락범위 제한 없음
		원자력발전 관련 엔지니어링용어	공공 데이터 포털	73건	용어정의, 영문설명, 한글설명	이용허락범위 제한 없음
		방사성물질 용어집	공공 데이터 포털	60건	용어, 용어설명	이용허락범위 제한 없음
		국내원전 약어집	공공 데이터 포털	3,453건	약어, 영문설명, 한글설명	이용허락범위 제한 없음
	한국국제교류재단	문화재용어사전	공공 데이터 포털	2,482건	시작자음(표제어), 한자, 독음, 설명	이용허락범위 제한 없음
	한국특허정보원	지식재산권용어	공공 데이터 포털	1,905건	용어(영문명), 설명, 출처	이용허락범위 제한 없음
	주택도시보증공사	보증업무 용어집	공공 데이터 포털	118건	용어, 설명	이용허락범위 제한 없음
	한국철도공사	철도 용어해설	공공 데이터 포털	1,135건	용어, 용어해설	이용허락범위 제한 없음
		철도통계 용어정의	공공 데이터 포털	724건	구분, 통계명, 용어명, 용어 설명	이용허락범위 제한 없음
	한국보건의료인국가시험원	시험 용어 해설	공공 데이터 포털	171건	대분류, 중분류, 용어, 영어표기, 정의	이용허락범위 제한 없음
	한국자산관리공사	선박금융 용어집	공공 데이터 포털	136건	용어, 설명	이용허락범위 제한 없음
	중소기업	중소기업	공공 데이터	24건	용어, 해설	이용허락범위

행정부처명	자료집명	자료 형태	건수	항목 정보	이용허락범위
기술정보진흥원	기술개발지원사업 용어정보	포털			제한 없음
한국환경산업기술원	환경 및 무역 관련 용어정의	공공 데이터 포털	30,831건	용어, 해설	이용허락범위 제한 없음
한국서부발전	발전용어집	공공 데이터 포털	5,063건	용어명	이용허락범위 제한 없음
한국환경공단	기후변화 용어	공공 데이터 포털	273건	한글명, 영문명, 설명	이용허락범위 제한 없음
	친환경차 종합정보 지원시스템 용어	공공 데이터 포털	17건	일련번호, 제목, 내용, 등록일시, 수정일시	이용허락범위 제한 없음
한국무역보험공사	무역보험 용어집	공공 데이터 포털	2,056건	용어명, 용어설명	이용허락범위 제한 없음
국가철도공사	철도 용어사전 관련 용어	기관 누리집	10,631건	용어명, 용어외래어 표기(영어, 한자), 용어 설명, 관련 용어	이용허락범위 제한 없음
대한석탄공사	석탄용어 사전	공공 데이터 포털	1,189건	용어명, 용어설명, 유사용어	이용허락범위 제한 없음
한국남부발전	해외사업 영문금융 용어	공공 데이터 포털	137건	영문용어, 뜻	이용허락범위 제한 없음
	발전용어집	공공 데이터 포털	5,064건	순서, 용어명, 한자명, 영문명	이용허락범위 제한 없음
한국지방재정공제회	지방예산 회계 용어집	공공 데이터 포털	314건	한글명, 영문명	출처 표시 + 변경금지
사립학교교직원연금공단	사립학교 교직원연금공단_용어사전	공공 데이터 포털	154건	순서, 한글단어이름, 영어단어이름, 한글내용,	이용허락범위 제한 없음

행정부처명		자료집명	자료 형태	건수	항목 정보	이용허락범위
					영어내용, 한글검색색인, 등록일, 등록자ID, 영어검색색인	
	대한무역투자진흥공사	해외투자 용어사전	공공 데이터 포털	602건	용어명, 영문명, 용어설명	이용허락범위 제한 없음
	한국국제협력단	ODA 용어사전	공공 데이터 포털	355건	한글명, 약어, 영문명	이용허락범위 제한 없음
	한국장학재단	공통포털_용어사전 기본	공공 데이터 포털	18건	용어명, 용어내용	이용허락범위 제한 없음
	서울특별시	행정데이터 표준용어 사전 정보	기관 누리집	535건	용어명, 설명, 영문약어명, 도메인명, 허용값, 저장 형식, 표현 형식, 행정표준코드명, 소관기관명	저작자 표시
합계				919,184건		

[표 7] 중앙행정기관 및 산하기관 공개 데이터베이스 현황

6. 공공기관 용어 자료 분석 종합

공공데이터 포털(data.or.kr)은 행정안전부 산하의 한국지능정보사회진흥원이 관리하는 누리집으로 다양한 공공데이터를 한곳에서 제공한다. 누구나 쉽고 편리하게 이용하도록 파일 데이터, 오픈 API 등 다양한 방식으로 제공하고 있고 누구든지 이용할 수 있도록 보장하며, 이용권의 보편적 확대를 위하여 노력하고 있다.(‘공공데이터법’ 제1조, 제3조)

공공데이터 이용허락 범위는 일반적으로 “이용허락범위 제한 없음”을 하고 있지만 일부 데이터의 대해서는 제한이 있으며 이에 대해 각 유형별로 저작권 표시를 참고하여 이용할 수 있다.

공공데이터 포털을 통해 용어 자료를 쉽게 가져다 사용할 수 있지만 JPG, PDF, HWP 등의 구조화 되지 않은 형식을 제공하는 경우가 있어 데이터를 제공하는 기관에서 구조화된 데이터를 제공받을 수 있는지 협의가 필요하다. 기관 자료마다 전체 항목과 항목명이 다르고 분야 간 정보가 연계되어 있지 않아 여러 분야에 속한 용어의 경우 분야별 용법을 찾아보기 힘들다. 또한 기관별 용어 자료 항목이 달라 사용자 입장에서 필요한 정보가 없거나 불필요한 정보가 제공될 수 있어 사전의 항목을 표준화할 필요가 있다. 기관이 제공하는 자료가 표준화되지 않고 정제되지 않은 경우도 있어 자료의 분석과 더불어 기관 협의를 통해 자료를 이관 및 정제하는 작업이 필수로 동반되어야 한다.

IV. 용어 통합 데이터베이스 연계 구축 방안

1. 용어 통합 데이터베이스 연계 구축 방안

용어 통합 데이터베이스 연계 구축을 아래 [그림 12]와 같이 4가지 방안으로 구분하였다. 현재 국립국어원에서 수행 중인 민관 합동 전문용어 총괄 지원 사업으로 마련되는 용어 목록을 시스템에 구축하는 방안, 다른 기관에서 사용 중인 용어 자료를 가져오는 방안, 다양한 종류의 언어 자료에서 용어를 발굴하는 방안, 기관에서 제공하는 용어 검색 오픈API를 연계하는 방안이 있다.

구분	용어 발굴 방안	내용 감수	분야 분류	전체 항목 기술	용어 표준화	특징	등급 (노출 여부)	통합 데이터베이스 연계 구축 방안
국립국어원 (민관 전문용어 지원단) 구축	• 매년 부처를 선정하여 전문용어 표준화 작업 (매년 약 200개) • 용어 발굴, 선정, 분야 분류, 정의문 작성, 대체 용어 마련, 표준화 작업	O	O	O	O	• 작업 품질 매우 좋음 ↑ • 매년 약 200건 정도의 적은 어휘 ↓	1 최상위 노출	• 매년 200개 어휘 구축
타 기관 용어 자료 가져오기	• 매년 2개 기관의 용어를 파일 혹은 데이터베이스 형태로 가져와 통합 데이터베이스에 이관 (총 10,000 어휘: 기관당 5,000개)	O	△	X	△	• 작업 품질 좋음 ↑ • 기관 보유 전문 용어를 일괄 등록하여 많은 어휘 ↑ • 표준화 미비 (어문 규범 감수 등) ↓	2 상위 노출	• 매년 2개 기관에서 어휘 발굴 • 분야 분류, 전체 항목 기술 등 1등급화 구축 작업 수행 (매년 8천건)
언어 자료에서 용어 발굴	• 말뭉치, 각 기관별 보도자료, 공개 가능한 문서, 웹문서 등 언어자료에서 딥러닝 기술 등을 활용하여 자동 발굴	X	X	X	X	• 대량의 언어 자원에서 반자동 발굴 하여 신속한 많은 어휘 발굴 장점 ↑ • 정의문도 없는 낮은 품질 ↓ • 선별, 중복 제거, 구축 등 보급을 위해 많은 작업 필요 ↓	4 노출안함	• 매년 5,000개 어휘 발굴 • 용어 발굴, 중복 제거, 선별을 거친 후 • 분류, 전체 항목 기술, 표준화 등 1등급화 구축 작업 수행
기관 용어 검색 오픈 API 연계	• 각 기관이 제공하는 검색 오픈 API를 활용 • 사용자가 보급 시스템에서 입력한 검색어에 따라 각 기관의 오픈 API에서 통합 검색된 용어 목록 획득	O	△	X	△	• 구축 부담 없이 다양한 전문 용어를 보급할 수 있음 ↑ • 오류 수정 등 관리는 각 기관이 담당하며 수정 사항이 바로 적용됨 ↑ • 표준화 미비 (어문 규범 감수 등) ↓	3 하위 노출	• 구축은 안함 • 보급 시스템에서 통합 검색으로 제공 (전문 용어 포탈 지향)

[그림 12] 용어 통합 데이터베이스 연계 구축 방안

첫 번째로 국립국어원에서는 매년 3~4개 부처를 선정하여 전문용어 표준화 민관 지원단을 구성하여 해당 부처 내에서 사용 중인 전문용어에 대해서 표준화 작업을 수행한다. 용어 발굴, 선정, 분야 분류, 정의문 작성, 대체용어 마련, 표준화 목록 확정의 절차를 거치며 각 단계별로 전문가들이 참여하여 표준화 작업을 수행하

게 된다. 그리고 국립국어원의 참여가 반드시 동반되기 때문에 용어의 전체 품질 역시 매우 좋다고 할 수 있다.

그 과정을 살펴보면 발굴 시에는 발굴하는 인력이 투입되어 각종 공문서 및 보도자료 등을 직접 살펴보면서 용어를 발굴하고 이 발굴된 용어를 전문가들의 참여를 통해 선정하는 과정을 거치며 다시 국립국어원의 검토를 거친다. 국립국어원의 검토를 거친 용어들을 전문가들에 의해 분야 분류, 정의문 작성, 대체용어 마련의 절차를 거치며 이러한 과정을 통해 표준안이 마련되면 표준화협의회, 국어심의회 등을 거쳐 표준 전문용어로 고시된다. 표준화 작업 시 각 단계별로 전문가들이 참여하여 서로 간의 의견을 여러 차례 주고받는 과정을 통해 전체적인 완성도를 높이고 있다.

다만 직접 사람이 발굴하고 선정하는 등의 일련의 과정을 거치게 되므로 긴 시간이 소요되고 표준화 가능한 용어가 200여 개로 매우 한정적이다. 사람 간 이루어지는 작업이므로 시간이 다소 소요되며 발굴자의 경우 일반인이 수행하는 경우 용어의 기본적인 항목을 기술하는 데 있어 상당한 어려움이 동반될 수 있다.

두 번째는 타 기관에서 사용되고 있는 용어 자료를 가져오는 방안이다. 한꺼번에 많은 기관의 용어 자료를 가져오는 것은 다소 부담이 될 수 있으므로 약 2개 기관을 선정하여 해당 기관에서 사용하는 용어 자료를 가져오는데 이때 가져오는 형태는 시스템에 즉시 적용할 수 있도록 파일 또는 데이터베이스 형태로 가져오는 것이 효율적이다. 이렇게 가져온 기관의 용어 자료를 모두 데이터베이스화하여 통합 데이터베이스를 구축하는 것을 목표로 한다.

이렇게 가져오는 용어 자료는 이미 사용 중인 자료이므로 실제적인 전문용어로서 효용성이 있으며 일괄 등록하게 되므로 많은 양의 용어 자료를 확보할 수 있다. 또한 이미 사용 중이기 때문에 어느 정도의 필요한 항목이 기술되어 있어 작업 품질이 좋다고 할 수 있다.

국립국어원에 직접 구축하는 용어의 수준으로 끌어올리기 위해서는 일관된 분야 분류 작업이 수반되어야 하고, 전체적인 내용 감수도 필요하다. 이러한 작업을 거치

게 되면 자료의 완성도가 더욱 높아질 것으로 기대할 수 있다.

세 번째 방안은 다양한 언어 자원들로부터 용어를 발굴하는 방안이다. 각 기관에서 생산한 보도 자료나 공개 가능한 공문서, 말뭉치, 웹 문서 등 다양한 언어 자료에서 딥러닝 기술 등을 활용하여 자동으로 전문 용어를 발굴하는 방안이다.

대량의 언어 자원으로부터 심층 학습 기술을 활용하여 자동 또는 반자동의 형태로 발굴하는 방안으로 사람의 개입이 적게 일어나므로 많은 양의 데이터 처리가 가능하기 때문에 짧은 시간 내에 많은 어휘를 발굴할 수 있는 장점이 있다. 그러나 기술만으로 용어 발굴은 가능하나 용어 자료로서의 항목 기술에 대한 부분이 미흡할 수 있다. 따라서 이러한 방법으로 발굴된 용어들의 품질을 확보기 위해서는 발굴 용어 중 중복 제거 등의 선별과정을 거치도록 하고 이후 보급을 위한 많은 작업들이 수반되어야 한다.

네 번째 방안은 타 기관에서 제공하고 있는 용어 검색 오픈 API를 연계하는 방안이다. 각 기관에서 제공하는 용어 검색 관련된 오픈 API들을 통해서 사용자가 검색하는 용어를 각 기관 오픈 API를 통해 검색하고 그 결괏값을 목록 형태로 획득하는 것으로 기관에서 이미 구축한 용어 사전을 쉽게 연계하여 사용할 수 있는 장점이 있다.

이렇게 오픈 API를 통한 연계 방안은 자체적으로 구축하지 않아도 이미 구축된 기관용어를 가져다 쓸 수 있어 구축에 대한 부담을 줄이고 오류사항을 각 기관의 담당자가 직접 수정하여 반영하므로 관리적인 측면에서도 부담을 줄일 수 있다.

직접적인 구축이 아니라 각 기관에서 제공하는 오픈 API를 통해서 연계하기 때문에 보급 시스템을 통해 검색할 수 있게 할 수 있으며 사용자에게 다양한 기관에서 보유하고 있는 용어 사전을 제공할 수 있다. 다만, 국립국어원 등 전문기관의 감수를 거치지 않았기 때문에 용어 표준화나 전체 항목 기술이 미흡할 수 있으므로 사용자에게 참고할 수 있는 용어로서의 노출이 적당하다고 할 수 있다.

이렇게 네 가지 방안으로 구축된 용어 통합 데이터베이스에서 사용자가 검색을

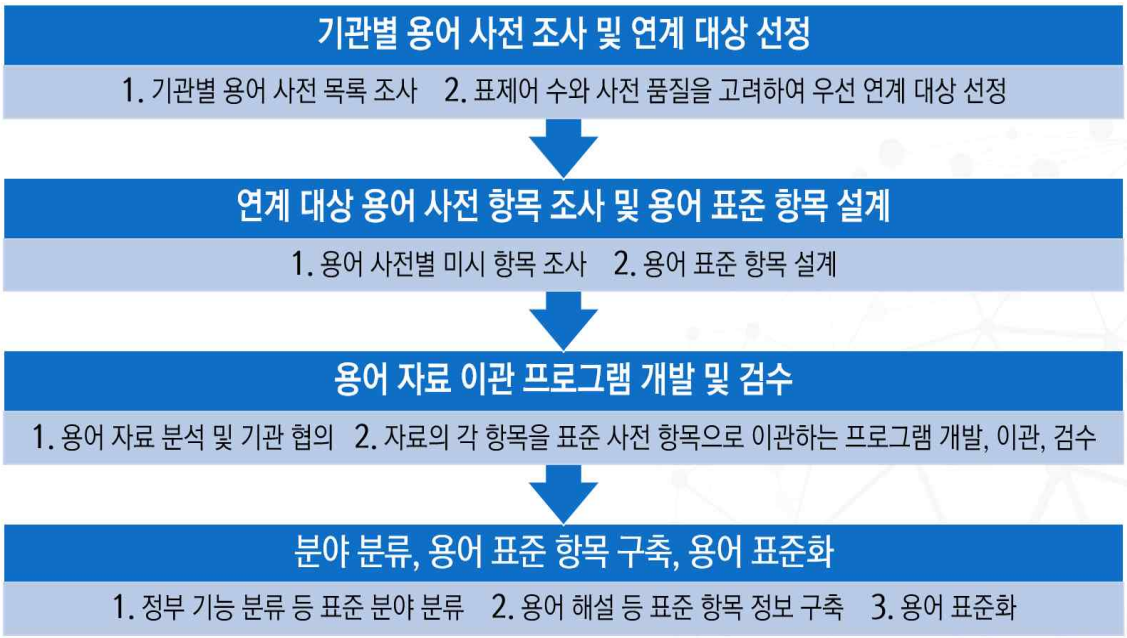
하면 가장 잘 표준화되어 있고 어문 규범에 맞게 정비 작업이 잘 이루어진 국립국어원에서 구축한 용어를 가장 먼저 노출하고 다음으로 타 기관에서 구축한 자료용어를 가져와 통합 데이터베이스에 이관한 자료를 두 번째로 노출한다. 그리고 기관 용어 검색 오픈 API를 통해서 가져온 결과를 세 번째로 노출한다. 그리고 언어 자원들로부터 자동으로 발굴한 용어들은 사용자들에게 노출하기에는 다소 미흡하므로 노출하지 않도록 한다. 발굴한 용어는 국립국어원의 주도로 용어 선별, 중복 제거, 항목 기술, 용어 표준화와 같은 작업을 수행하도록 한다.

2. 기관별 용어 자료 연계 구축 방안

기관의 용어 자료를 모아 연계 구축하기 위해서는 연계할 대상의 기관들을 선정하는 것이 첫 번째로 수행되어야 한다. 해당 기관에서 연계할 수 있는 용어 사전의 종류는 어떤 것들이 있는지 용어 사전을 조사하여 사전의 항목은 어떻게 되는지 표제어 수는 몇 개인지 등의 전체적인 품질을 검토한다. 조사된 항목 정보에 따라 구축해야 할 용어 표준 항목의 이관 방법을 설계한다. 데이터베이스 통합 구축을 위해 용어 사전 이관 프로그램을 개발하게 되는데 이때 연계해야 할 용어 자료의 구성 및 자료 형식 등에 관해서 해당 기관과 협의하여 방안을 구성한다. 자료 구성 및 자료 형식이 확정되면 용어 자료 이관을 하는 프로그램을 개발하고 개발된 프로그램을 이용하여 기관의 용어 자료 항목을 표준 항목으로 이관하게 된다. 이렇게 이관된 항목에 대해 협의가 이뤄진 사항이 모두 반영되었는지를 검수하게 된다.

이렇게 이관된 타 기관의 용어 자료를 과학 기술 표준 분류 체계, 정부 기능 분류 체계, 우리말샘 분류 체계에 맞게 분류하며 표준 항목인 표제어, 용어 설명, 용례, 분야, 원어, 대체 용어 등의 항목 중에서 작성되어 있지 않은 항목에 대해서는 정보를 추가로 구축하도록 한다. 또한 같은 개념이 여러 용어로 쓰여 혼란을 일으킨다면 하나로 통일하고, 어려운 용어는 이해하기 쉬운 용어로 어문 규범에 맞도록 정비하는 작업을 함께 수행하여 국립국어원에서 제공하는 고품질의 용어 데이터베

이스의 수준에 맞출 수 있도록 한다. 이러한 과정을 아래 [그림 13]으로 도식화하였다.



[그림 13] 기관별 공공용어 자료 연계 구축 방안

3. 용어 표준 항목 설계 방안

국립국어원 내에서 구축하는 용어 데이터베이스의 일관성을 확보하기 위해서는 우선적으로 용어 표준 항목 설계가 이루어져야 한다. 용어 표준 항목 설계를 통해 향후 수행되는 타 기관의 용어 자료 연계 시 타 기관에서 구축된 용어 자료에 추가적으로 정보 구축이 필요한 항목이 무엇인지 파악할 수 있고 구성된 표준 항목에 따라서 이관될 수 있도록 기준점을 제시할 수 있다. 표준 항목은 사용자의 이해를 돕고 용어 자료로서의 가치를 가지도록 분야, 대상 용어, 원어, 표준안(대체 용어), 용어 해설(정의문), 사용 예시 등 6개 항목을 제안한다.

원어와 용어 해설(정의문)은 대상 용어를 이해하는 데 꼭 필요한 정보이고, 사용 예시는 대상 용어를 활용하는 데 도움을 줄 수 있다. 국어심의회 등을 통해 마련된 표준안(대체 용어)을 제시하여 용어 사용자들이 쉽고 편리하게 사용할 수 있도록 정보를 제공한다. 분야 분류는 「과학기술기본법」에 따라 만들어진 국가 과학기술 표준 분류 체계를 따르도록 하여 기관, 학계, 산업계에서 활용할 수 있는 기반을 다진다.

분야		대상 용어	원어	표준안 (대체 용어)	용어 해설 (정의문)	대상 용어 사용 예시
대분류	중분류					
보건의료	치료기기 | 치료/진단기기	엠아르아이	MRI(magnetic resonance imaging(영))	자기^공명^영상	강한 자기장 내에서 인체에 라디오파를 전사해서 반향되는 전자기파를 측정하여 영상을 얻어 질병을 진단하는 검사.	11월부터 복부·흉부 MRI(→자기공명 영상) 검사비 부담 3분의 1로 줄어든다.

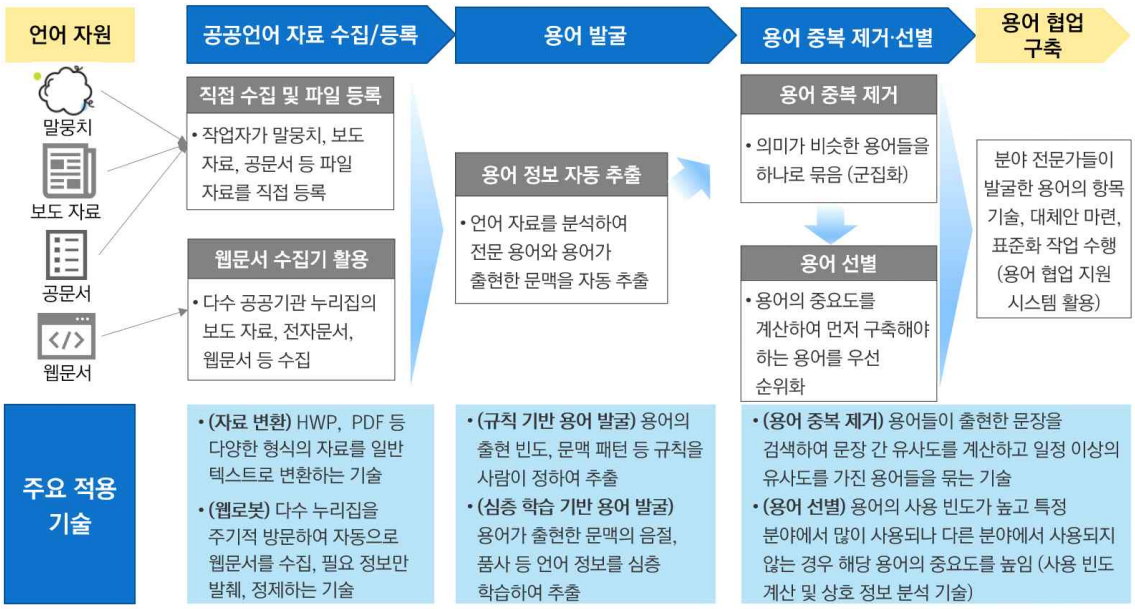
[그림 14] 용어 표준 항목 설계와 작성 예시

이렇게 용어 표준 항목을 설계하게 되면 이에 따라 용어 데이터베이스를 구축하게 되고 타 기관 용어 자료 연계나 이관 시에도 설계된 항목에 따라 구축이 이루어질 수 있다. 따라서 항목을 설계할 때에는 향후 구축되어야 할 기관들의 용어들이 통합 데이터베이스에 구축될 수 있도록 항목을 설계하는 것이 중요하다.

V. 언어 자원 기반 체계적 용어 발굴, 분석, 표준화 방안

1. 개요

전문용어 등 모든 공공용어를 분야별로 체계적으로 구축하려면 전문 분야별 말뭉치 구축이 필요하다. 이러한 말뭉치를 확보하기 전이라도 단기적으로는 새로운 공공용어 발굴이나 관련 정보를 추출하는 데 참고할 수 있도록 기존의 언어 자료를 수집하여 활용하는 것이 필요하다. 즉 보도 자료와 같은 공문서, 웹문서 등 다양한 언어 자료를 직접 수집하거나 웹문서 수집기 등을 통하여 자동으로 수집해서 언어 자원을 구축할 필요가 있다. 이렇게 구축한 언어 자원으로부터 용어 후보를 자동으로 발굴한 후 중복 제거와 선별 작업을 통해서 우선 표준화가 되어야 할 용어를 선정한다. 이렇게 선정된 용어를 용어 협업 구축 지원 시스템을 활용하여 각 분야의 전문가들이 표준화된 용어 항목을 기술하고 대체안을 마련하며 표준화 작업을 수행한다.



[그림 15] 언어 자원 기반 체계적 용어 발굴, 분석, 표준화 방안

2. 용어 발굴 방안

2.1. 용어 자동 발굴 방안

전문용어 자동 추출에 대한 연구[14]는 용어의 품사 패턴, 용어의 어휘 패턴, 정의문에 대한 구문 패턴 등 수동 구축 패턴을 이용한다.

구문특징패턴	형식특징패턴	항목특징패턴
N 은/는 ~이다.	N :	용어(의) 정의
N 은/는 ~으로(서/써)	- N	
N 은/는 ~말하-	* N :	
N 은/는 ~로(서/써)	N -	
N 은/는 ~ 정의/총칭/지칭/용어		
N 라/란 ~을/를 말하-		
N 은/는 ~이고/이며		
N 은/는 ~을/를 의미하-		
N 은/는 ~ㅁ		
N 은/는 ~라(고) 할 수 있-		
N 라/란 ~으로(서/써)		
~을/를 N 라 하-		
N 라/란 ~이다.		
N 라/란 ~용어		
N 은/는 ~을/를 뜻하-		
N 라/란 ~을/를 의미하-		
N 라/란 ~라(고) 하-		

[그림 16] 용어 추출을 위한 정의문 패턴의 예

[그림 16][15]와 같이 용어가 출현하는 구문의 특징 패턴을 미리 구축한 후 문서가 입력되면 주어진 패턴을 매칭하여 전문용어를 추출하는 것이다. 이와 같은 방법은 패턴을 수동으로 구축하는 데 따른 초기 비용이 많이 들고, 수동으로 구축된 패턴에 의존적이라는 단점이 있다.

이런 단점을 보완하기 위해 심층 학습 기반으로 용어를 추출하는 방법이 있다.

14) 신효식 외(2002). 텍스트로부터 용어 정의문의 자동 추출 방법. 한국정보과학회 언어공학연구회 / 한희정 외(2017). 기술문서 정의문 패턴을 이용한 전문용어사전 자동추출 및 활용방안. 정보관리학회지

15) 한희정 외(2017). 기술문서 정의문 패턴을 이용한 전문용어사전 자동추출 및 활용방안. 정보관리학회지

심층 학습은 사람이 수동으로 패턴, 규칙을 구축하는 것이 아니라 학습 데이터를 통해 자동으로 규칙을 생성하고 생성된 규칙을 통해 용어를 추출하는 방식이다.

일례로 '단어 임베딩'(word embedding)을 이용하는 방법이 있다. 단어 임베딩은 단어를 수치화해 사용하는 방법의 하나로서 단어를 공간상의 벡터로 표현하는 기술이다. 문서에 있는 모든 단어를 대상으로 단어 임베딩 기술[16]을 적용하여 만든 벡터의 유사도를 이용하여 전문용어 여부를 판단하는 것이다. 이 방법은 전문용어는 유사한 단어가 적을 것이라는 가정에서 출발한다. 전문용어는 희소성 특성이 있으므로 유사성이 일반 단어보다 현저하게 적다고 판단하여 추출하는 방식이다.

2.2. 용어 수동 발굴 방안

전문용어를 발굴하기 위해 기술적인 요건도 중요하지만 각 정부부처 및 산하기관, 과학기술단체 등에서 보유하고 있는 산재된 다양한 용어 사전 및 용어 데이터베이스, 말뭉치를 취합하고 이를 기반으로 하여 취합된 목록을 내에서 사람이 검토하여 수동으로 용어를 발굴할 수 있다.

그뿐만 아니라 현재 실질적으로 업무를 수행하고 있는 전문용어 민관합동 총괄지원단이 용어를 발굴하는 방식과 같이 정부부처 및 산하기관의 보도 자료 및 신문기사, 공문서 등 다양한 문서에서 직접 발굴하는 방법이 있으며 기계가 찾아내지 못하는 용어를 사람의 눈으로 확인하고 발굴하는 장점도 있다.

3. 용어 중복 제거 방안

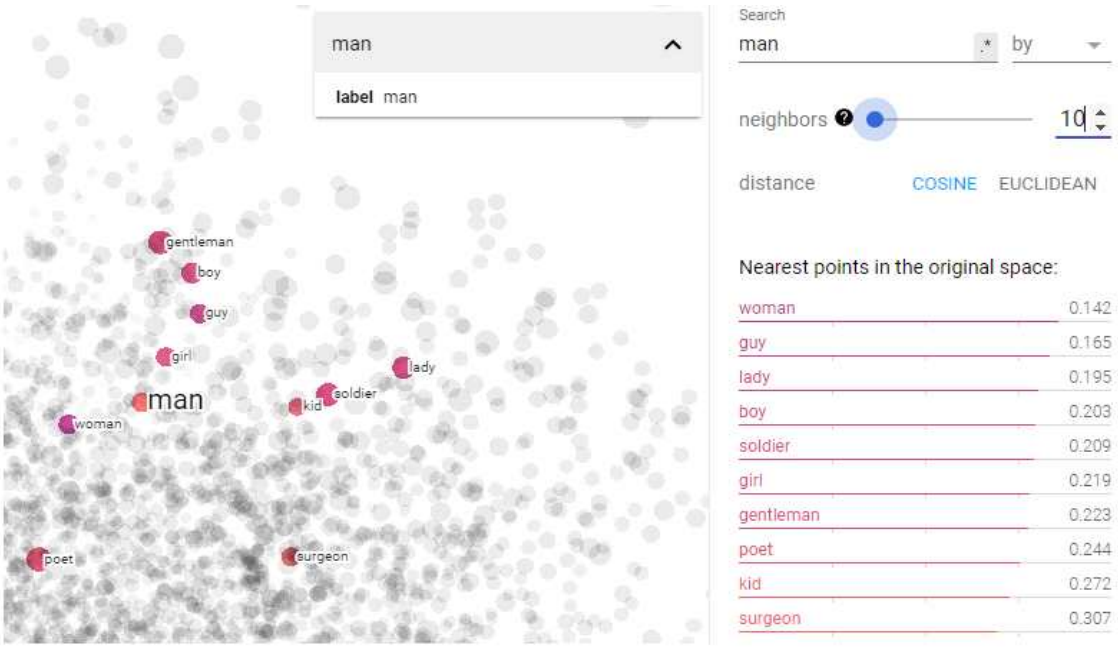
용어 중복 제거는 의미가 비슷한 용어를 하나로 묶는 작업이다. 용어 중복 제거를 자동화하는 방안으로 임베딩 벡터를 이용하는 방안과 검색 기술을 이용하는 방안을 제안한다.

임베딩 벡터를 이용하는 방안은 모든 용어의 임베딩 벡터를 구한 후 공간에 사상하여 계층적 클러스터링(hierarchical clustering)[17]을 수행하면 비슷한 의미의 용

16) 'word2vec', 'FastText', 'GloVe' 등의 기술이 있다.
17) 계층적 클러스터링은 개체들을 가까운 집단부터 계층적으로 묶어나가는 방식으로 군집 개수를 정하지 않아도 군집분석이 가능하다.

어가 군집화된다. 군집화된 용어를 시각화하여 제시하면 작업자가 대표 용어를 결정하여 유사 용어를 묶는 작업을 수행한다. 이 방식은 성능 좋은 심층 학습 모델을 활용하게 되면 검색 기술 이용 방식에 비해 정확도가 상대적으로 높다는 것이 장점이다. 단점은 심층 학습 서버 등 고가의 자원이 필요하며 계산량이 많아 규칙을 만드는 학습 과정이 느리다는 것이다.

검색 기술을 이용해서 비슷한 용어를 군집화할 수 있다. 용어가 출현한 문장으로 검색하여 문장 간 유사도를 계산하고 일정 기준 이상의 유사도를 가진 문장을 선별하면 비슷한 용어가 군집화된다. 이 중에서 대표 용어를 결정하고 유사 용어를 묶는 작업을 수행한다. 이 방식의 장점은 심층 학습보다 계산량이 월등히 적고 단순하므로 추출 대상 언어 자원이 매우 크지만, 서버 자원을 적게 사용하고 속도가 빠르다는 것이다. 단점은 검색 엔진의 유사도 측정 방식에 따라 군집화 정확도가 결정되며 유사도 측정을 하지 않는 검색 엔진은 적용하기 어렵다.



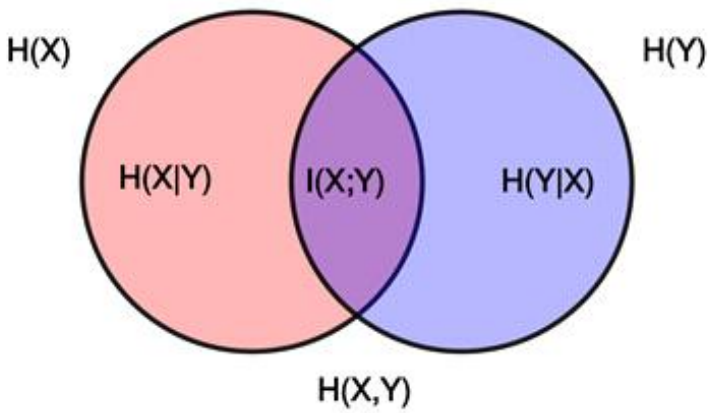
[그림 17] 임베딩 벡터에 의한 군집 시각화 예시 (출처: projector.tensorflow.org)

4. 용어 선별 방안

용어 선별은 용어의 중요도를 계산하여 먼저 구축해야 하는 용어를 순위화하는 것을 말한다. 앞서 언급한 것처럼 언어 자원으로부터 심층 학습 기술 또는 규칙 기반 기술을 적용하여 용어를 자동으로 추출하면 많은 수의 용어 후보를 추출하게 된다. 시간, 인력, 비용 등 주어진 자원에 한계가 있으므로 많은 후보 용어 중 어떤 용어를 먼저 구축해야 하는지 순위를 정할 필요가 있다.

여러 문서에서 많이 출현한 용어가 중요도가 높다는 가정에 용어의 출현 빈도를 계산하여 순위를 정하는 방법은 가장 간단하며 효율적이라는 장점이 있다. 단점은 용어의 출현 빈도만 고려하면 용어의 전문 분야별 특성이 반영되지 못한다는 것이다.

이에 용어의 출현 빈도뿐만 아니라 상호 정보를 함께 고려하여 순위를 정하는 방법을 제안한다. 상호 정보(mutual information)는 두 확률 변수가 서로 어떤 관계를 맺고 있는지 나타내는 정보를 말한다. [그림 18]은 상호 정보를 벤다이어그램으로 개념화한 것으로 두 확률 변수 H(X)와 H(Y)의 교집합인 I(X;Y)가 상호 정보량에 해당한다. 두 확률 변수 H(X), H(Y)가 독립이라면 교집합이 존재하지 않으므로 상호 정보량이 0이 된다.



[그림 18] 상호 정보의 개념

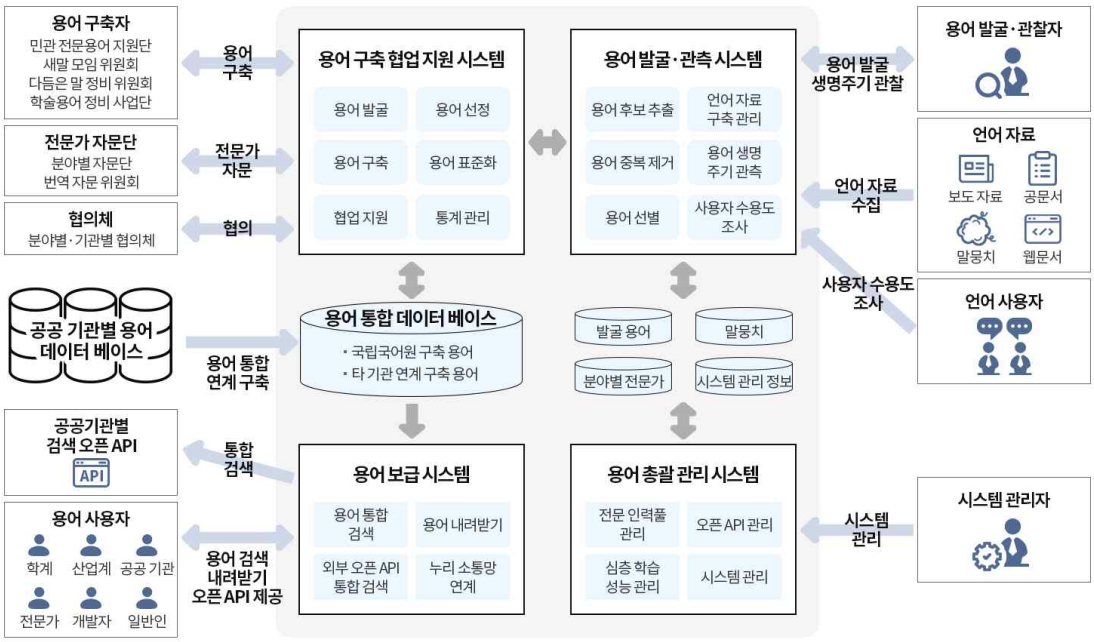
용어의 우선 순위화 관점에서 상호 정보를 설명하면 출현 빈도가 높은 특정 용어가 전문 분야 A와 상호 정보량이 많고 다른 전문 분야 B에서는 상호 정보량이

적다면 특정 용어는 전문 분야 A에서 우선순위가 높다고 판단한다. 반대로 출현 빈도가 높은 특정 용어가 전문 분야 A, B 모두 상호 정보량이 많다면 전문 용어가 아닌 일반 용어로 판단할 수 있다. 간단히 설명하면 출현 빈도와 상호 정보를 고려한 용어의 중요도 계산은 특정 용어가 해당 분야에서만 사용되고 다른 분야에서 사용되지 않을 경우 해당 분야에서 중요하다고 판단하는 것이다.

VI. 용어 구축, 관리, 관측, 보급 시스템 구축 방안

1. 목표 시스템 개념도

향후 구축되어야 할 용어 구축, 관리, 관측, 보급 시스템의 전체 목표 시스템을 개념도를 작성하였다. [그림 19]와 같이 용어 구축 협업 지원 시스템, 용어 보급 시스템, 용어 발굴·관측 시스템, 용어 총괄 관리 시스템 등 4개의 시스템으로 구성되어 있다.



[그림 19] 목표 시스템 개념도

용어 구축 협업 지원 시스템에서는 발굴한 용어 후보 중 구축할 용어를 선정하고 대상 용어, 용어 해설(정의문), 분야, 원어, 사용 예시 등 표준 항목에 따라 용어 정보를 작성한 후 동일 의미의 여러 용어를 통일, 어려운 용어를 쉬운 용어로 다듬기, 어문 규범에 맞게 정비하는 등의 표준화 작업을 수행한다. 이러한 과정에 여러 작업자가 참여하게 된다. 용어 구축은 민관 전문용어 지원단, 새말 모임 위원회, 다듬은 말 정비 위원회, 학술용어 정비 사업단에 맡는다. 전문 용어의 정확한 기술을 위해 전문 분야별 자문단, 번역 자문 위원회에 자문을 의뢰한다. 전문 용어를 담당

하는 분야별, 기관별 협의체와 용어 선정, 표준화 등을 협의하게 된다.

용어 보급 시스템은 학계, 산업계, 공공 기관, 분야별 전문가, 개발자, 일반인 등 다양한 유형의 사용자가 용어를 쉽고 편하게 활용할 수 있도록 보급하는 시스템이다. 용어를 검색하는 기능, 용어 목록을 MS 엑셀 등 파일로 내려받는 기능 등 용어를 검색·활용·유통하는 기능을 제공한다.

용어 보급 시스템에서 사용자에게 다양하고 풍부한 용어를 보급하고 용어 구축 협업 지원 시스템에서 구축할 용어를 발굴하기 위해 여러 공공 기관에서 보유한 용어 데이터베이스를 본 시스템에 등록하여 용어 통합 데이터베이스를 구축한다. 국립국어원이 보유한 기존 용어 사전을 등록하고 타 기관의 용어를 오픈 API, 데이터베이스, 파일 등의 여러 연계 방식에 활용하여 용어 통합 데이터베이스를 구축한다.

용어 발굴·관측 시스템에서는 용어 구축 협업 지원 시스템에서 구축할 용어 후보를 보도 자료, 공문서, 말뭉치, 웹문서 등 언어 자원으로부터 자동으로 발굴한다. 자동으로 발굴한 대량의 용어 중 작업할 용어를 추리기 위해 비슷한 의미의 용어를 중복 제거하고 용어의 중요도를 계산하여 우선순위를 매기고 작업할 용어를 선별한다. 용어 발굴을 위한 언어 자료를 구축하기 위해 언어 자료를 구축 관리 하는 기능을 제공한다. 용어 구축, 표준화 과정을 거처 보급한 용어가 사회에서 널리 사용되고 정착하는지 관측하는 기능도 제공한다. 용어가 생성되고 사라질 때까지 수명주기를 엿볼 수 있는 용어 생명주기 관측 기능을 제공한다. 보급한 용어를 사용자가 얼마나 수용하고 있는지를 설문 등을 통하여 조사하는 기능을 제공한다.

용어 총괄 관리시스템은 앞서 언급한 세 개의 시스템을 총괄 관리하는 시스템이다. 분류 정보, 코드 정보 등 세 개 시스템의 공통 정보를 구축 관리 하는 기능을 제공한다. 그 외 분야별 전문가 인력 풀 관리, 심층 학습 성능 관리 등 전체 시스템에 관련된 기능을 제공한다. 이처럼 용어 총괄 관리 시스템은 시스템 관리자만 접근할 수 있는 기능과 정보로 한정하여 엄격한 접근 제한을 통해 보안 문제에 대응한다.

2. 대구 정부통합전산센터 클라우드 기반 아키텍처 설계

국립국어원은 2022년도에 완공되는 대구 정부통합전산센터 입주 대상으로 선정되어 있으므로 환경에 맞게 아키텍처를 구성해야 한다. 대구 정부통합전산센터에서 제시하는 아키텍처 기본 설계 방향은 인터넷망, 업무망(행정기관), 업무망(공공기관) 등으로 영역을 분리해야 하며 Web, WAS, DB를 분리하여 3개의 계층 구조로 설계하고 이중화하여야 한다. 또한, 업무 단위로 시스템을 분리하여 증설 및 확장에 용이하도록 구성해야 한다.

위 설계 방향에 맞추어 용어 구축 협업 지원 시스템, 용어 보급 시스템, 용어 발굴·관측 시스템, 용어 총괄 관리 시스템 등 4개의 업무 단위로 시스템을 분리하였다. 용어 구축 협업 지원 시스템, 용어 보급 시스템은 용어 구축을 위한 외부 전문가와 용어 사용자 등 일반 사용자가 사용해야 하므로 인터넷망에 위치해야 한다. 용어 발굴·관측 시스템, 용어 총괄 관리 시스템은 업무(공공기관)에 위치한다. 이러한 점을 고려하여 아래와 같이 아키텍처를 설계하였다.

용어 구축 협업 지원 시스템

구분	하드웨어	스토리지 (GB)	대수	OS	Web	Was	그 외 SW	비고
Web 서버	소형	100	2	RHEL	Apache	-		• 외부 전문가 협업을 위해 인터넷망에 구성
WAS 서버	중소형	100	2	RHEL	-	Jboss		
DB 서버	중소형	300	2	RHEL	-	Cubrid		

용어 발굴·관측 시스템

구분	하드웨어	스토리지 (GB)	대수	OS	Web	Was	그 외 SW	비고
Web 서버	소형	100	2	RHEL	Apache	-		
WAS 서버	중소형	1024	2	RHEL	-	Jboss		• 대용량의 언어 자료를 저장하기 위해 1TB 스토리지 필요
DB 서버	중소형	100	2	RHEL	-	Cubrid		
웹수집 서버	소형	100	1	RHEL	-	Jboss	웹로봇	
검색서버	소형	100	2	RHEL	-	JBoss	검색엔진	
심층 학습 서버	별도 서버*	300	1	RHEL	-	Jboss	Python, Tensorflow	• 심층 학습 전용 서버 구성 • 용어 구축 협업 지원 시스템과 심층 학습 서버 공유

용어 보급 시스템

구분	하드웨어	스토리지 (GB)	대수	OS	Web	Was	그 외 SW	비고
Web 서버	중소형	100	2	RHEL	Apache	-		
WAS 서버	중형	100	2	RHEL	-	Jboss		
DB 서버	중형	100	2	RHEL	-	Cubrid		
검색서버	중형	100	2	RHEL	-	JBoss	검색엔진	

용어 총괄 관리 시스템

구분	하드웨어	스토리지 (GB)	대수	OS	Web	Was	그 외 SW	비고
Web 서버	소형	100	2	RHEL	Apache	-		
WAS 서버	중소형	100	2	RHEL	-	Jboss		
DB 서버	중소형	100	2	RHEL	-	Cubrid		

[그림 20] 대구 정부통합전산센터 기반 아키텍처 설계

3. 연도별 시스템 구축 방안

연도별로 용어 구축 협업 지원 시스템, 용어 발굴·관측 시스템, 용어 보급 시스템, 용어 총괄 관리 시스템의 상세 구성 방안 계획을 수립하였다.

	2021년	2022년	2023년
용어 구축 협업 지원 시스템	용어 협업 구축 기틀 마련 • 용어 구축 및 협업 기능 • 용어 통계 및 작업 통계 기능 • 용어 구축 지침 조회 및 검색 기능	일반 사용자 참여형 용어 구축 • 개방형 용어 구축 협업 기능 • 개방형 용어 구축 관리 기능 • 누리 소통망 활용 토론 연계 기능	자연어 처리 기술 기반 용어 구축 • 전문 분야 자동 분류 기능 • 용례 후보 자동 추출 기능
용어 발굴 · 관측 시스템	용어 발굴·관측 기반 마련 • 언어 자료 구축 관리 기능 • 용어 후보 추출 기능 • 용어 생명주기 관측 기능	체계적인 용어 발굴·관측 기능 구현 • 용어 중복 제거 기능 • 용어 선별 기능 • 용어 생명주기 통계·분석 기능	사용자 수용도 조사 기능 구현 • 사용자 설문 기능 • 누리 소통망 수집 및 분석 기능 • 미등록 검색어 수집 및 분석 기능
용어 보급 시스템	공공 용어 포털 구축 • 용어 통합 검색 기능 • 외부 오픈 API 통합 검색 기능 • 용어 내려받기 기능	개인화 맞춤화 기능 구현 • 전문 분야별 맞춤형 용어 서비스 기능 • 개인 맞춤형 용어 서비스 기능 • 누리소통망 연계 기능	전문가를 위한 고급 활용 체계 구축 • 공공 용어 오픈 API 서비스 기능 • 전문 용어 자동 인식 기능 • 용어 생명 주기 서비스 기능
용어 총괄 관리 시스템	시스템 총괄 관리 체계 마련 • 전체 시스템 공유 자료 관리 기능 • 전문 인력풀 관리 기능 • 외부 오픈 API 관리 기능	용어 보급·맞춤화 관리 체계 마련 • 공공 용어 오픈 API 관리 기능 • 전문 분야별 맞춤화 서비스 관리 기능 • 개인 맞춤형 용어 서비스 관리 기능	성능 분석·강화 시스템 구축 • 심층 학습 기반 도구 학습 관리 기능 • 심층 학습 기반 학습 자료 관리 기능

[그림 21] 연도별 시스템 구축 계획

2021년에는 용어 구축 협업 시스템은 용어 협업을 위한 기틀을 마련한다. 용어 구축 및 협업 기능을 구현하고 용어 및 작업 통계, 용어 구축 지침 조회와 검색 기능을 구현한다. 용어 발굴·관측 시스템에서는 용어를 신규 발굴하기 위한 언어 자료 구축 및 관리 기능, 언어 자료에서 용어 후보를 추출하는 기능, 용어의 생명주기 관측 기능을 구현한다. 용어 보급 시스템에서는 공공용어 포털 구축을 위해 용어 통합 검색 기능과 외부에 오픈 API를 통해 검색할 수 있는 기능, 용어를 파일로 내려받는 기능을 구현한다. 용어 총괄 관리 시스템은 전체 시스템에서 공유되는 자료 관리 기능과 용어 구축 협업 지원 시스템에서 같이 협업 업무를 수행할 전문 인력풀을 관리하는 기능, 외부 오픈 API를 관리하는 기능을 구현한다.

2022년에는 용어 구축 협업 지원 시스템을 일반 사용자들에게도 개방하여 참여할 수 있는 기능을 구축한다. 또한 개방형 용어 구축 관리와 누리 소통망을 활용한 토론 연계 기능을 구축한다. 용어 발굴·관측 시스템에서는 자료에서 뽑아낸 용어들

가운데서 중복을 제거하는 기능과 중복 제거를 통해 용어를 선별하는 기능, 용어 생명주기 관측에서 한발 더 나아가 통계 및 분석하는 기능을 구축한다. 용어 보급 시스템에서는 전문 분야별 맞춤형 서비스를 제공하는 기능을 구축하고 개인 맞춤형 용어 서비스 기능을 구축한다. 또한 누리 소통망 연계로 용어를 편리하게 공유하는 기능을 구현한다. 용어 총괄 관리 시스템에서는 용어 보급 시스템에서 제공하는 오픈 API 서비스를 관리하는 기능을 구현하고 맞춤화된 용어 보급 서비스를 관리하기 위한 전문 분야별 맞춤화 서비스, 개인 맞춤형 서비스 관리 기능을 구현한다.

2023년에는 용어 구축 협업 지원 시스템에 자연어 처리 기술을 적용하여 좀 더 편리하게 용어를 구축할 수 있는 기능을 구현한다. 전문 분야를 자동으로 분류하는 기능, 용례 후보를 자동으로 추출하는 기능을 구현한다. 용어 발굴·관측 시스템에서는 설문 기능과 누리 소통망의 수집 및 분석 기능, 미등록 검색어 수집 및 분석 기능을 구현하여 용어의 사회적 수용도를 직접 또는 간접적으로 조사할 수 있도록 한다. 용어 보급 시스템에서는 전문가를 위한 공공용어 오픈 API를 서비스하고 전문 용어 자동 인식 기능을 구현한다. 용어 생명주기 관측 서비스를 외부에 개방하여 전문가가 활용할 수 있도록 한다. 용어 총괄 관리 시스템에서는 심층 학습 도구 학습 관리 기능과 학습 자료 관리 기능을 구축한다.

4. 연도별 소요예산

시스템 구축과 용어 데이터베이스를 구축하기 위한 연도별 수행되어야 할 상세 내역에 맞추어 연도별 예산 계획을 수립하였다. 아래 [표 8]과 같이 총 3단계로 구분하여 크게 공공용어 구축과 구축, 보급, 관리 시스템 구축으로 구분하고 공공용어 구축 내에서도 기관별 기구축 자료 정비 및 DB 연계와 공공언어 자료 구축하는 것으로 구분하였다. 시스템 구축에서는 용어 구축 협업 지원 시스템, 용어 발굴·관측 시스템, 용어 보급 시스템, 용어 총괄 관리 시스템 구축 비용과 하드웨어 및 소프트웨어 도입을 위한 비용으로 구분하여 예산 계획을 수립하였다.

구분			1단계			2단계	3단계
			’21년	’22년	’23년	’24년~’26년	’27년~’29년
공공 용어 구축	기관별 기 구축 자료 정비 및 DB 연계	기관 용어 사전연계 구축	200 (50백만 원 *4기관)	200 (50백만 원 *4기관)	200 (50백만 원 *4기관)	300 (50백만 원 *2기관/연)	300 (50백만 원 *2기관/연)
		공공용어 목록 구축·정비 용역	400 (20,000원 *2만건/연)	800 (20,000원 *4만 건/연)	800 (20,000원 *4만 건/연)	1,200 (20,000원 *2만 건/연)	1,200 (20,000원 *2만 건/연)
구축, 보급 관리 시스템 구축	공공언어 자료 구축 검색		150	150	150	450	450
	용어 구축 협업 지원 시스템		100	150	150	450 (유지보수, 고도화: 1.5억 원/연)	450 (유지보수, 고도화: 1.5억 원/연)
	용어 발굴·관측 시스템		150	130	150		
	용어 보급 시스템		150	130	150		
	용어 총괄 관리 시스템		100	100	150		
	하드웨어 및 소프트웨어 구입		70	30	30		
합계			**1,320**	**1,690**	**1,780**	**2,400**	**2,400**

[표 8] 연도별 소요 예산(안)

공공용어 구축은 용어 통합 데이터베이스를 구축하는 것이다. 다른 기관이 보유한 용어 자료를 연계해서 구축하고 정비하는 것과 용어를 발굴하기 위해 공공언어 자료를 구축하는 것으로 나누었다.

기관별 기구축 자료 정비 및 DB 연계에서는 기관 용어 사전을 연계하여 구축하는 비용과 공공용어 목록을 구축하고 정비하는 용역비로 구분하였다. 기관 용어 사전 연계 구축은 1단계에서는 매년 4개 기관을 대상으로 하고 기관당 5천만 원의 예산을 책정했다. 2단계 및 3단계 시에도 매년 2개 기관씩 확장해 나가는 계획을 수립하였다. 공공용어 목록 구축·정비 용역에서는 연계된 기관 용어 데이터베이스의 각 항목을 과학 기술 표준 분류 체계에 따라 분류하고 용어의 원어 정보를 정비하고 용어의 개념을 이해하고 활용할 수 있도록 정의문과 용례를 정리하는 등의 정비에 인력이 투입되며 건당 20,000원으로 계산하여 2021년에는 2만 건, 2022년과 2023년에 각각 4만 건의 공공용어를 정비한다. 2단계 및 3단계에서도 계속해서 연

20,000개의 용어를 구축 및 정비를 해서 지식을 표상하는 용어의 보고를 확충하도록 하여야 한다.

공문서 언어 자료 구축은 국립국어원이 보유한 공문서, 감수 자료와 다른 기관이 보유한 공공 자료를 언어 자료를 구축하는 비용으로 매년 1억 5천만 원의 예산을 책정했다. 이후에도 지속적으로 자료를 구축해 나갈 수 있도록 매년 1억 원의 예산 계획을 수립했다.

구축, 보급, 관리 시스템 구축은 용어 구축 협업 지원 시스템, 용어 발굴·발굴 관측 시스템, 용어 보급 시스템, 용어 총괄 관리 시스템 등 4개 시스템에 대한 3년 구축 예산을 책정했다. 이후 2단계와 3단계 시에는 유지보수 및 기능 고도화 같은 사업에 연 1.5억 원의 예산을 책정했다. 또한 2022년 대구 정부통합전산센터에 이전하기 전에 필요한 하드웨어와 검색 솔루션, 웹 수집 솔루션 등을 고려하여 소프트웨어 구입 예산을 책정했다.

담당 연구원 강미영(국립국어원 공공언어과 학예연구관)
이현주(국립국어원 공공언어과 학예연구사)

사업 책임자 이영현(㈜엔에이치엔다이퀘스트)
사업 참여자 윤철진(㈜엔에이치엔다이퀘스트)
이수정(㈜엔에이치엔다이퀘스트)
이성훈(㈜엔에이치엔다이퀘스트)
최지영(㈜엔에이치엔다이퀘스트)
박새나(㈜엔에이치엔다이퀘스트)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2020년 12월 12일
발행일: 2020년 12월 12일
인 쇄: 세광제록스