

2022 인공지능학습용데이터 구축사업

과제발굴을 위한 수요조사 계획

목 차

Ⅰ. 인공지능 데이터 구축 사업 설명

Ⅱ. 인공지능 학습용 데이터 수요조사 개요

Ⅲ. 수요조사 양식 및 과제 예시

Ⅰ. 인공지능 학습용 데이터 구축 사업 설명

01 사업 개요

배경 한국판 뉴딜 종합계획(20년 7월)의 “데이터 댐” 프로젝트의 핵심 사업

➔ DNA 생태계를 통한 체질 전환과 일자리 창출을 위해
“인공지능 학습용 데이터“ 1,300종 구축(25년까지)

개요 데이터(이미지, 동영상, 텍스트, 음성 등)를 **수집(획득)**하여 기계학습 개발에 필요한 형태로 **가공**하여 **개방**(AI허브 : www.aihub.or.kr)하는 사업

범위 ① 데이터 수집(획득), ② 데이터 가공, ③ AI 모델(알고리즘) 개발

영역 음성 | 자연어 | 비전 | 안전 | 환경 | 헬스케어 | 자율주행 ……

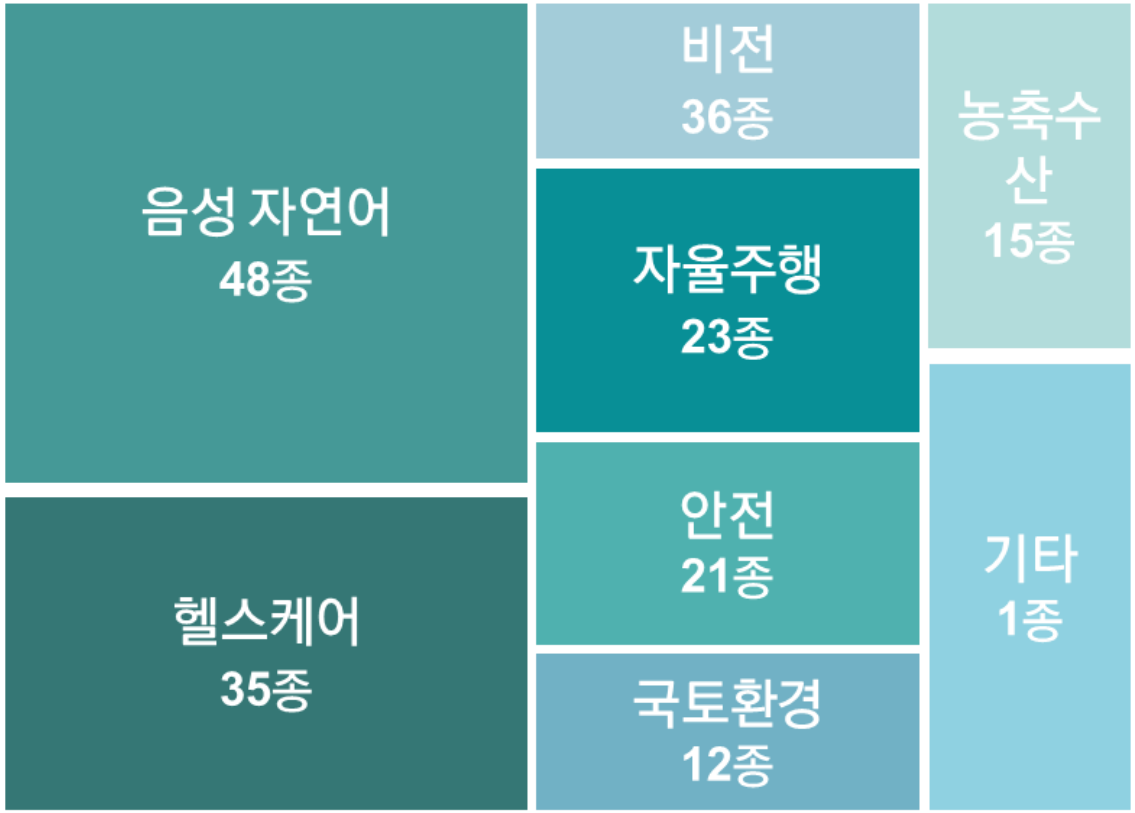
주무부처 과학기술정보통신부 **전담기관** 한국지능정보사회진흥원(NIA)

02 사업 현황

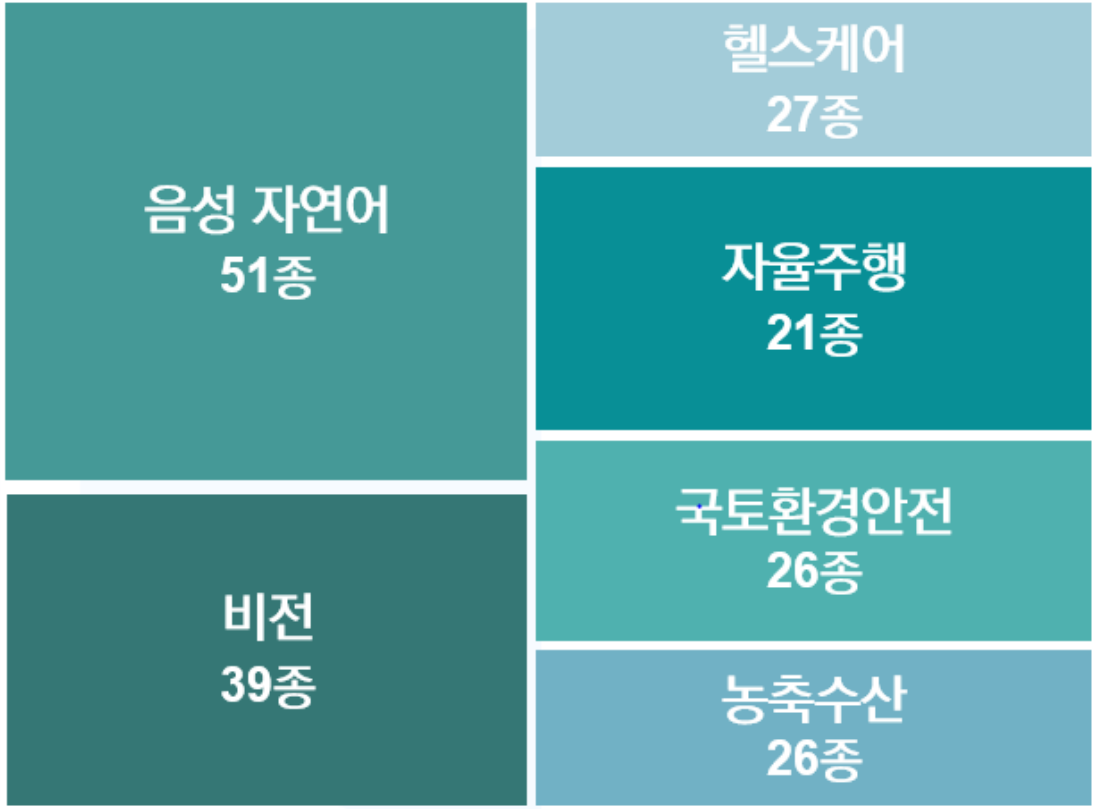
전문 분야와 범용성이 높은 인공지능 학습용 데이터를 구축 하여

현재 AI허브(www.aihub.or.kr)에 개방 ('18년 : 4종, '19년: 7종, '19년 12월: 10종, ' 20년: 170종)

개방 '17~20년(총 191종)



'22년 개방 현재 구축 중 '21년(190종)



03 구축 데이터 현황

법률
교통사고, 퇴직금 등 생활관련 법령, 판례, 사례, 법률용어, 생활용어 등의 관계정보를 구조화한 지식 데이터(21만건)

특허
국내 전기전자, 기계화학분야에 출원·등록된 특허의 특허정보, 심사정보, 특허기술용어 등의 관계정보를 구조화한 지식 데이터(170만건)

한국형 이미지(음식/자동차)
150여종의 한국 음식 이미지 데이터 및 100종의 한국 차량 이미지 데이터 (20만장)

일반 상식
한국어 위키백과에서 추출한 객체정보와 관련정보의 관계를 구조화한 지식 데이터(15만건)

관광
국내 주요 관광지의 식당, 시설, 표지판 등을 촬영한 이미지와 다국어 POI 정보를 결합한 데이터(150만장)

농업
국내 농업관련 영농정보, 상담정보, 지원사업정보 등의 관계를 구조화한 지식 데이터와 농작물 2종의 병충해 이미지 데이터(5만건)

질병진단(안저 이미지)
녹내장, 당뇨망막증 등 주요 안과질환의 안저 촬영 이미지와 전문의 소견을 결합한 데이터(7천5백장)

한국어 음성
2~3명이 서로 자연스럽게 연속하여 발화하는 한국어 음성과 발화 내용을 전사한 데이터(1,000시간)

한국어 대화
음식점, 소매점, 학원 등 소상공인 업종에서 자주 발생하는 상황과 질문·답변·의도 등을 구조화한 대화 데이터(50만건)

멀티 모달(드라마)
영상 내에 사람의 감정, 행동, 상황, 대화 내용 등의 멀티모달정보를 결합시킨 동영상 데이터(20시간)

기계 독해
다양한 질문과 정답, 정답을 포함하고 있는 지문의 쌍으로 구성된 지능형 QA용 텍스트 데이터(45만건)

질병진단(유방암 엑스레이 이미지)
국내에서 유병률이 높은 유방암 촬영 엑스레이 이미지 및 결과 데이터 (3.5만장)

03 구축 데이터 현황

한국어-영어 번역 말뭉치
한국어 인공지능 번역 기술개발 및 성능강화를 위한 한국어-영어 병렬 말뭉치 데이터 160만 문장

한국형 사물 이미지
한국형 강화를 위한 사물/거리 등 이미지 데이터 360만장
객체, 장소, 상황 인지기술 개발 및 성능

한국어 글자체 이미지
광학글자인식(OCR) 성능개선을 위한
한글 글자체(손글씨 및 인쇄체) 이미지 660만자

인도보행 영상
보행지원기술 개발을 위한 인도 보행 및 인도 위
객체(사람, 자전거, 펜스 등) 라벨링 데이터 67만건

멀티모달(융합) 영상
감성인식 AI개발을 위해 동영상에서 인물의 표정, 음성, 발화내용, 상황 등의 감정이 포함된 멀티모달 영상 데이터 100시간

사람동작 영상
사람의 동작·자세·행동 인식기술 개발을 위해 다양한 조건에서 사람 동작 영상 데이터 20만 건과 영상에서 추출한 관절정보 이미지 400만장

한국인 안면 이미지
총 1,000명에 대해 여러 각도, 조도 등의 환경 하에서 안면인식·식별 성능강화를 위한 한국인 얼굴 데이터 1,900만장

위해물품 엑스레이 이미지
위험물·도구 자동판별 기술개발 및 성능개선 위한
범죄 도구, 반입금지물품 등의 X-ray 이미지 48만장

이상행동 CCTV 영상
주야간 적용가능한 이상행동 지능형 탐지기술개발 및
성능 강화를 위한 CCTV 영상 데이터 700시간

AI 학습용 데이터 360종 구축 개방

사업기간	2022년 3월 ~ 12월 31일 * 사업기간 추후 변동 가능
사업규모	약 360종 AI 학습용 데이터 구축 및 개방(aihub.or.kr) *데이터 종수 추후 변동 가능
지원대상	인공지능 학습용 데이터를 구축·개방하고자 하는 기업, 대학, 공공기관, 지자체 등
사업목적	양질의 AI 학습용 데이터 구축·개방을 통한 AI 응용서비스·제품 개발 활성화와 AI의 기존 산업계 적용 · 확산 지원, 데이터 전문가 육성으로 일자리 창출
주요내용	① AI 데이터 기획 → ② AI 데이터 수집 및 가공 → ③AI 데이터 검증 및 개선 → ④ AI 데이터 학습모델 개발 → ⑤ AI 데이터 공개 및 활용, 확산

[참고] AI허브 : www.aihub.or.kr



Ⅱ. 인공지능 학습용 데이터 수요조사 계획

수요조사 개요

목적 2022년도 신규 과제(데이터) 발굴 및 기획을 위한 수요조사

조사 내용 인공지능 학습용 데이터 수요
데이터 분야, 타입, 활용 분야 등 자유롭게 제안

조사 기간 21.9.15(수) ~ 9.30(목)

제출방법 네이버폼 http://naver.me/GMRrVpOt 통해 설문 작성 또는 첨부한 수요조사 설문지를 작성하여 이메일(airoadmap@nia.or.kr) 회신

문의처 NIA 인공지능데이터전략팀 (02-6747-2169, 2166), airoadmap@nia.or.kr

Ⅲ. 인공지능 학습용 데이터 수요조사 양식

01 인공지능 학습용 데이터 수요설문조사지 설명

활용 영역	인공지능 활용 영역 선택
활용 목적	인공지능 활용 목적(서비스, 솔루션, AI 임무, 기능 등) 상세하게 기술
데이터명	필요한 인공지능 학습용 데이터명 기술
데이터 개요	위에서 제시한 데이터를 일반인도 이해할 수 있는 수준으로 설명
데이터 형식	데이터 형식 선택(이미지, 영상, 음성, 텍스트 등)
데이터 수집 장비	원천 데이터 수집장비를 다음 중 선택(스마트폰, 마이크, 카메라, 센서 등)
데이터 수량	앞서 제시하신 서비스, 기능 등 구현에 필요한 필수 데이터 수량을 상세하게 기술(예시 : 이미지 ##만장)
보유 데이터 현황	데이터 보유 여부

참고 과제 예시

데이터명	주요 영역별 회의 음성데이터
데이터 개요	한국어로된 회의영상/음성을 인식하여 자동으로 회의록을 작성하고 자막을 생성을 위한 회의 음성 데이터 수집
데이터 구축량	도메인별 회의(실제환경, 시사토론, 독서모임, 온라인 회의, 라디오프로그램등) 관련 음성 데이터 7,000시간 이상
활용 분야	회의록 자동 작성 및 요약 서비스, 대면/비대면 회의 통번역 서비스등
원시 데이터 형태	음성 데이터 및 전사 텍스트
원시 데이터 수집 디바이스	스마트폰, 노트북, 컨퍼런스용 마이크 등
원시 데이터 수집 방안	온라인 컨퍼런스 영상 및 스마트폰 , 노트북 마이크를 통해 특정 주제에 대한 자유 녹음 및 시나리오 기반 녹음/ 다양한 화자의 목소리가 겹치는 상황 등 참여자가 서로 균등하게 이야기하는 구성 방안 제시
데이터 가공형태	일관된 원칙으로 음성을 전사 (transcription), 영문/숫자 한글로 표시, 영문/숫자를 그대로표시, 발화 앞뒤 묵음 500~1000ms 포함

Q&A

문의 : NIA 인공지능데이터전략팀 (02-6747-2169, 2166)
airoadmap@nia.or.kr