

더 많은 데이터 줄게, 더 똑똑해져라! 인공지능 학습용 데이터, 역대 최대 규모로 개방한다.

- '22년도 구축 데이터 310종, 오늘부터 7월말까지 매달 약 80여종씩 순차개방
- 기존 6대 분야에서 제조·로보틱스 등 14대 분야 691종으로 데이터 다양화
 - AI허브(aihub.or.kr)에 회원가입만 하면 국민 누구나 활용할 수 있어

과학기술정보통신부(장관 이종호, 이하 '과기정통부')와 한국지능정보사회진흥원(원장 황종성, 이하 '지능정보원')은 '22년도에 「인공지능 학습용 데이터 구축 사업」을 통해 구축한 학습용 데이터셋 310종을 오늘부터 7월말까지 'AI허브(aihub.or.kr)'를 통해 순차적으로 개방한다고 밝혔다.

동 사업은 전문영역부터 일상생활까지 전분야로 확산되고 있는 인공지능 기술개발과 지능화 서비스 확산을 선도하기 위한 국가 핵심 데이터 인프라 사업이다. 기업, 연구자, 개발자는 물론 인공지능 개발 등에 관심 있는 우리 국민이라면 누구나 'AI허브'를 통해 학습용 데이터를 활용할 수 있다.

과기정통부와 지능정보원은 2020년부터 본격적으로 민간·공공 수요 등에 맞춰 매년 약 2백 종의 데이터를 구축해 왔다. 전년도 구축 데이터는 당해연도 품질검증 등을 거쳐 개방하는데, 지금까지 개방된 데이터는 한국어, 이미지·영상, 헬스케어 등 6대 분야 381종, 약 11억 건에 이른다. 지난해 7월에는 190종의 데이터를 개방하면서, 'AI허브' 연간 방문자 수가 최초로 1백만 명을 돌파했다.

올해 개방하는 데이터는 310종으로 역대 최대 규모이다. 지난해 구축 분야를 6대 분야에서 제조·로보틱스·교육·금융·스포츠 등의 14대 분야로 확장하면서

사업 규모를 크게 늘렸기 때문이다. 이번에 310종 약 15억 건의 데이터 개방을 완료하면 'AI허브' 이용자들은 총 691종, 약 26억 건의 데이터를 활용할 수 있게 된다. 특히 올해는 대규모 개방인 점을 고려하여, 이용자들이 더 많은 데이터를 조금이라도 빨리 활용할 수 있도록 품질검증 등을 먼저 마친 데이터를 우선 개방한다. 이에 따라 초기 개방 시점은 지난해 보다 3개월 앞당겨진다. 매달 약 80여 종 내외를 개방하여 7월말까지 전체 310종의 개방을 완료할 계획이다.

이번 1차 개방에는 자연어 기반 질의·검색·생성 데이터, 고서(古書)한자·한글 등 광학인식(OCR) 데이터, 안전환경 조성을 위한 실내외 군중 특성 데이터, 소방대원 행동 모션 3차원(3D) 객체 데이터 등 자연어와 인공지능 비전 분야에서 광범위하게 활용될 수 있는 약 70종의 데이터가 포함되어 있다.

한편 'AI허브'에 공개되는 데이터들은 국제 표준에 맞는 품질기준을 달성하고 개인정보 비식별화를 거친 경우에 한하며, 기업·기관들이 실제 사용하고 있는 인공지능 모델에 직접 데이터를 학습시켜 유효성, 정확성 등을 확인하는 활용성 검토를 받는다. 또한 개방 이후 약 3개월 동안은 사용자들로부터 데이터 품질 요구사항, 오류 등의 의견을 반영하여 데이터 품질 제고를 위한 보완 과정을 거칠 예정이다.

엄열 인공지능정책관은 “챗GPT 등 초거대 인공지능의 등장으로 인공지능 산업이 빠르게 발전하고 있다.”며 “우리 기업들과 연구자들이 새로운 데이터 확보에 어려움을 겪지 않도록 기존의 라벨링 데이터(지도학습용) 중심의 사업을 개편하여 초거대 인공지능 학습에 필요한 대규모 비라벨링 데이터와 한번에 여러가지 유형을 학습시킬 수 있는 다중임무형 라벨링 데이터 등 다양한 유형의 데이터를 지속 확보해 나갈 예정”이라고 밝혔다.

담당 부서	인공지능정책관	책임자	팀 장	최우석 (044-202-6280)
	인공지능확산팀	담당자	사무관	박수정 (044-202-6281)

