보도일시 :
2022. 1. 19.(수) 석간, 온라인 1. 19 10:00부터

디지털 뉴딜의 성공을 위한 대표 프로젝트,

인공지능(AI) 학습용 데이터 구축 · 활용 고도화 방안

2022. 1. 19.

관계부처 합동

요 약

인공지능 학습용 데이터 구축 · 활용 고도화 방안(요약)

Ⅰ 인공지능 학습용 데이터 구축 추진배경

□ 정부는 '**인공지능 국가전략**'('19.12), '**디지털 뉴딜**'('20.7)을 통한 디지털 전환 가속에 대비하며 대표 프로젝트로 '**데이터 댐**' **구축** 본격화('20.9~)

○ 전 산업·사회적인 **인공지능 도입·확산**의 **핵심자원**이 될 **인공지능 학습용 데이터 구축·개방**에 **'25년**까지 **2.5조원**의 **대규모 투자** 계획

※ '17~'19년 21종 → '25년까지 총 1,300여종 구축·개방 추진

< (참고) 인공지능 학습용 데이터 구축의 필요성 >

◇ **인공지능 모델**의 **데이터 학습량**은 **성능과 비례**하며, 각 분야로 인공지능 기술이 확산·발전되기 위해서는 **분야별 고품질·대규모 인공지능 학습용 데이터 확보**가 필수

◇ 단, **데이터 수집·가공**에 **시간·비용 소요***가 커서 중소·스타트업, 대학 등의 **인공지능 도입·확산**에 **장벽**이 되므로, **국내 실정에 맞는 데이터**의 **양적·질적 확충** 요구 증대

* 국내 AI·데이터 기업은 AI 개발 시간의 80%, 비용의 75%가 데이터 확보에 소요된다고 응답('20, NIA)

➡ 지속 제기되는 **데이터 부족 문제 해소**와 **디지털 전환 가속화**를 위한 **체계적인 데이터 자원 확보·활용**을 뒷받침할 **전략방안 마련** 필요

Ⅱ 데이터 구축 · 개방 현황 및 주요성과

□ **[현황] 기존 21종**('17~'19)에 더하여 **'20년 구축**한 **170종 4.8억건***의 데이터를 **'21.6월 개방**하였고, **'21년 190종 추가 구축**(잠정 '22.上 개방)

* 다양한 수요분석과 전문가 검토를 통해 음성·자연어, 비전, 헬스케어 등 8개 분야 선정·구축

□ **[주요성과]** 데이터 추가개방 후 약 6개월간 **6만여 건**의 **다운로드**와 함께, 이를 활용한 다양한 **인공지능 혁신서비스 창출**

< AI Hub 데이터를 활용한 혁신서비스 창출 사례 예시 >

라이드플럭스	· **자율주행 데이터** 기반 **무인이동 서비스 도입·고도화** 및 **지역 확대** ※ 제주관광지 자율주행 셔틀('21.8), 세종정부청사 승객이동 서비스('21.9) 운영
아워랩	· **수면 질 분석 데이터**를 활용한 **수면자세 감응형 수면무호흡 치료기기 개발** ※ 식품의약품안전처 의료기기 품목허가 취득('21.2)
씨유박스	· **개인동의 기반 안면이미지 데이터**를 활용한 **위변조 감지 알고리즘** 개발 ※ 국내 기업 최초 관련 기술 국제표준(ISO/IEC 30107-3) 인증 획득('21.7)

○ '20년 추경 등을 통한 **데이터 구축 과정**에 기업·대학·병원 등 **571개 기관**과 청년·경력단절여성 등 **4만여명 이상**의 **인력 참여**

※ 데이터 품질 강화를 위해 데이터 구축 전과정에 활용할 수 있는 품질기준도 마련·적용('20.10~)

Ⅲ 고도화 방안 추진방향 및 비전

□ **[추진방향] 충분한 과제수요***와 **구축·활용 주체**들의 **개선·보완 의견**을 토대로, 1,300종 데이터의 **전략적 구축방향 설정**과 **활용 활성화 유도**

* 디지털 뉴딜 추진 후 산·학·연 각 계 수요조사를 통해 3천3백종 이상의 데이터 수요 확보

□ **[비전 · 목표 및 과제]**

비전

풍부한 데이터 자원 기반, 인공지능 기술·산업 선도국 도약

목표

	2021	2022	2023	2024	2025
데이터 구축	누적 340종		900종	1,100종	1,300종+α
데이터 다운로드(활용)	누적 25만건		70만건		100만건
인공지능 개발·확산	누적 3천건		8천건		1만건

과제

1. 데이터 구축·개방 로드맵 수립·이행

❶ 로드맵의 방향성과 전략을 제시합니다.
❷ 시장이 원하고 국가가 필요로 하는 데이터, 이렇게 구축됩니다.
❸ 로드맵 이행체계를 갖추어 나갑니다.

2. 구축 데이터의 활용성과 가치 증진 지원

❶ 쉽게 접근하고, 편리하게 활용할 수 있도록 합니다.
❷ 다양한 정부 지원사업과 연계하여 활용가치를 높여 갑니다.
❸ 믿고 쓸 수 있는 양질의 데이터로 만들어 갑니다.
❹ 데이터 구축에 함께한 분들의 성장을 지원합니다.

Ⅳ 추진과제

1 데이터 구축 · 개방 로드맵 수립 · 이행

① 체계적 데이터 확보를 위한 방향과 전략을 제시합니다

"기반기술 + 2대 전략분야를 축으로 종적 · 횡적 확장"

① (기반기술) **한국어, 영상·이미지** 등으로 **여러 기술·산업**의 **공통기술**로 활용되는 **범용적 성격**의 **인공지능* 구현**을 지원하는 분야

* 시각·청각·인식·동작 등을 비롯하여 다양한 분야에서 활용될 수 있는 공통 기술

② (전략분야1) 헬스케어·교통·안전 등 **공익 증진**에 기여할 수 있고, **조기에 인공지능 확산**과 **성과 창출**을 **기대**할 수 있는 분야

③ (전략분야2) 인공지능을 통한 **산업·영역별 혁신**을 **촉진**할 수 있는 분야

※ 로드맵 수립을 위한 정책연구 추진 및 전문가, 데이터 활용기업, 개발자 등 의견 지속 수렴('20.10~)

2 시장이 원하고 국가가 필요로 하는 데이터, 이렇게 구축됩니다

구 분		목 표	구축 방향	
			단기(~'23)	중기(~'25)
기반 기술	한국어	다양한 분야·상황에 적용될 한국어 인공지능 개발 지원	단일한 환경, 특정 조건하의 음성·대화·질의응답·번역 등 기초데이터 확보	저음질·다화자 음성, 다양한 목적·상황별 복합 정보로 고도화
	영상·이미지	포괄적 영상·이미지 확보로 관련 기술 연구 선도	객체·공간 인식을 위한 다양한 형태·행동·환경 데이터 확보	대규모 복합 상황 및 공간 인식을 위한 복합 정보로 고도화
전략 분야 1	헬스케어	인공지능 기반 질병진단·치료 기기·서비스 개발 지원	질병판독·진단 지원, 수면·건강관리 및 임산부 등 기초데이터 확보	질병 재발·악화 등 예측, 의료환경(병원안전) 및 소외계층 데이터 등 확장
	교통·물류	작동이 불필요한 수준의 자율주행, 스마트 물류 실현	다양한 도로환경 객체, 교통상황, 차량 내외부 및 기초 물류데이터 확보	각종 이상상황, 도로 외 공간 객체·상황과 물류과정 전주기 데이터로 확대
	재난·안전·환경	인공지능 기반 재난·재해예측, 선제적 대응체계 마련	개별재난, 고위험 분야 인지 및 환경(피복·산림 등) 기초데이터 확보	복합재난, 각종 생활 안전사고 예측·인지 및 기후변화 데이터 등 고도화
	농·축·수산	자율형 스마트 농·축·수산 서비스 개발과 확산	국내 주요 작물 및 축종·어종별 생육·행동 기초데이터 확보	생산, 질병관리 및 유통·소비 등 전주기 데이터로 확대
전략 분야 2	제 조	공정 최적화, 지능형 제조솔루션 개발을 위한 공정·품질 정보 등		
	로보틱스	용도·산업별 로봇 기술 개발·고도화를 위한 인지·모션 데이터 등		
	법 률	법률 지식·접근성 증진, 서비스 지능화를 위한 법령·용어·판례 등		
	교 육	맞춤형·지능형 교육을 위한 분야별 교재, 학습분석 정보 등		
	금 융	정보 비대칭 해소, 서비스 개발·고도화를 위한 투자분석정보, 용어 등		
	지식재산	특허상담, 유망기술 예측을 위한 특허·기술·분쟁·거래 정보 등		
	문화·관광	지능형 문화·관광 서비스를 위한 이미지, 지식DB 등		
	스포츠	경기력 향상, 전술 고도화 등을 지원할 영상·이미지, 각종 규칙정보 등		
	➡ 상기 예시분야 외 인공지능 도입·확산을 위해 전략분야 지속 발굴·확장			

☞ 매년 변화하는 정책환경을 반영하여 세부 추진계획 보완 및 차년도 구축과제 기획·선정 추진

3 로드맵 이행 체계를 갖추어 나갑니다

○ **주요 영역별 전문성**을 보유한 **부처·전문기관**이 **기획위원회 참여**를 통해 **과제기획, 추진방향 설정** 등 구축·활용 전반을 **주도**

❶ 사례1 : 농축산	해당 분야 전문성을 갖춘 **농식품부**가 **기획-구축-활용**을 **주관**하는 방식 도입·적용('22) 후 **수요**에 따라 **분야 확산**
❷ 사례2 : 한국어	**과기정통부-문체부(국립국어원)** 간 상호 **전략적인 데이터 기획-구축**을 위한 **협력체계** 운영, **데이터 연계** 등 추진

2 구축 데이터의 활용성과 가치 증진 지원

1 쉽게 접근하고, 편리하게 활용할 수 있도록 합니다

○ **(클라우드 토탈서비스) 활용빈도**가 높은 일부 **데이터**를 다운로드 없이 **클라우드 상**에서 **가공, 컴퓨팅자원 이용** 등이 가능토록 지원('22~)

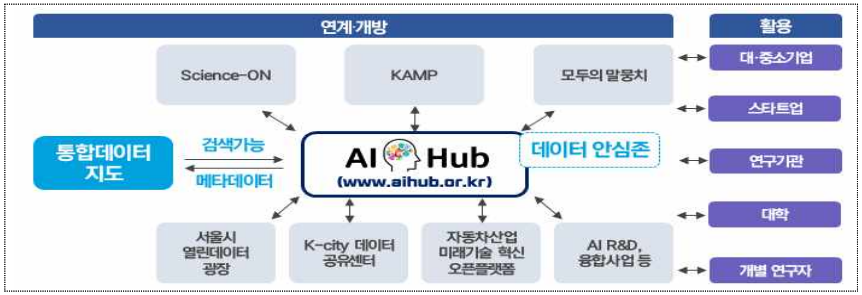
※ 장기적으로 AI Hub(스토리지)를 광주 데이터센터로 이관, 컴퓨팅자원 등 원스톱 제공('23~)

○ **(데이터 접근성 개선) 이용자 편의 증진**을 위한 **개선방안** 시범운영

데이터 분할제공	대용량의 데이터를 **필요한 만큼**만 내려 받아 **활용**할 수 있도록 **소규모·중규모**(1~20%) **버전 데이터** 제공 병행('22~)
객체 선택제공	**이용자가 원하는 객체만 선택**하여 다운로드 할 수 있도록 데이터에 포함된 **개별 객체를 검색**하여 **부분 다운로드** 제공('22~)
데이터 딜리버리	네트워크 환경이 열악한 이용자도 **대규모 데이터**를 자유롭게 활용할 수 있도록 **저장매체 기반 '주문-대여 서비스'** 제공('22~, 자율주행 분야 시범)

○ **(데이터 활용·환류)** 구축된 **데이터**를 **서비스 개발**에 활용하면서 **파생된 데이터**를 **AI Hub**에 **축적·개방**하는 **시범서비스** 운영('22~)

○ **(AI Hub 연계·확장)** AI Hub 데이터가 자유롭게 **수정·재가공**되어 확산되도록 **기업 플랫폼 게시 허용** 및 **결과물 AI Hub 공유** 추진('22~)



○ **(AI Hub 기능 고도화)** 핵심데이터 중심으로 **상시 경진대회(리더보드)**를 운영하고, **교육·실습자료 탑재** 및 **커뮤니티 기능 강화**('22~)

○ **(개인정보보호 강화) 주기적 데이터 검증**과 함께 **'개인정보 신고제'**를 도입하고('22~), 유형별 **가명·익명처리 가이드라인**을 마련·적용('22~)

※ '개인정보 보호활용 기술 R&D 로드맵에 따라, 데이터 유형별 개인정보 탐지 등 기술개발 및 활용 지원('22~)

2 다양한 정부 지원사업과 연계하여 활용 가치를 높여갑니다

○ **(D.N.A. 지원사업간 연계) 디지털 뉴딜** 등을 통해 추진 중인 다양한 **데이터-클라우드-AI 지원사업**을 연계한 **패키지형 지원** 추진('22~)

< 연계 지원 체계도(안) > * (예시) A기업, B기업

○ **(정부 지능정보화사업 연계)** 과기정통부 내 **지능정보화사업** 등과 **우선 연계**하고, 향후 **타 부처 지능정보화사업**과 **연계 확대** 추진('22~)

○ **(타 사업←데이터 제공) 인공지능 R&D, AI+X** 등 **인공지능 융합사업** 추진 시 **AI Hub 데이터 활용 확대 지원** 및 **일부 의무화** 검토('22)

3 믿고 쓸 수 있는 양질의 데이터로 만들어 갑니다

○ **(선제적 품질관리) 데이터 구축 중간단계**부터 **인공지능 모델학습**을 **병행**하고, 오류 점검·보완 등을 위한 **사전검증** 및 **자문** 실시('22~)

○ **(기존 데이터 고도화) 추가·보완**이 필요한 데이터를 **발굴·고도화**하는 **'기 구축 데이터 고도화' 과제 유형 신설** 추진('22)

○ **(맞춤형 데이터 구축)** 데이터 성격과 구축 난이도 등을 고려해 **지원규모, 구축기간**을 **다양화**하고 **별도 운영·평가체계 마련·적용**('22~)

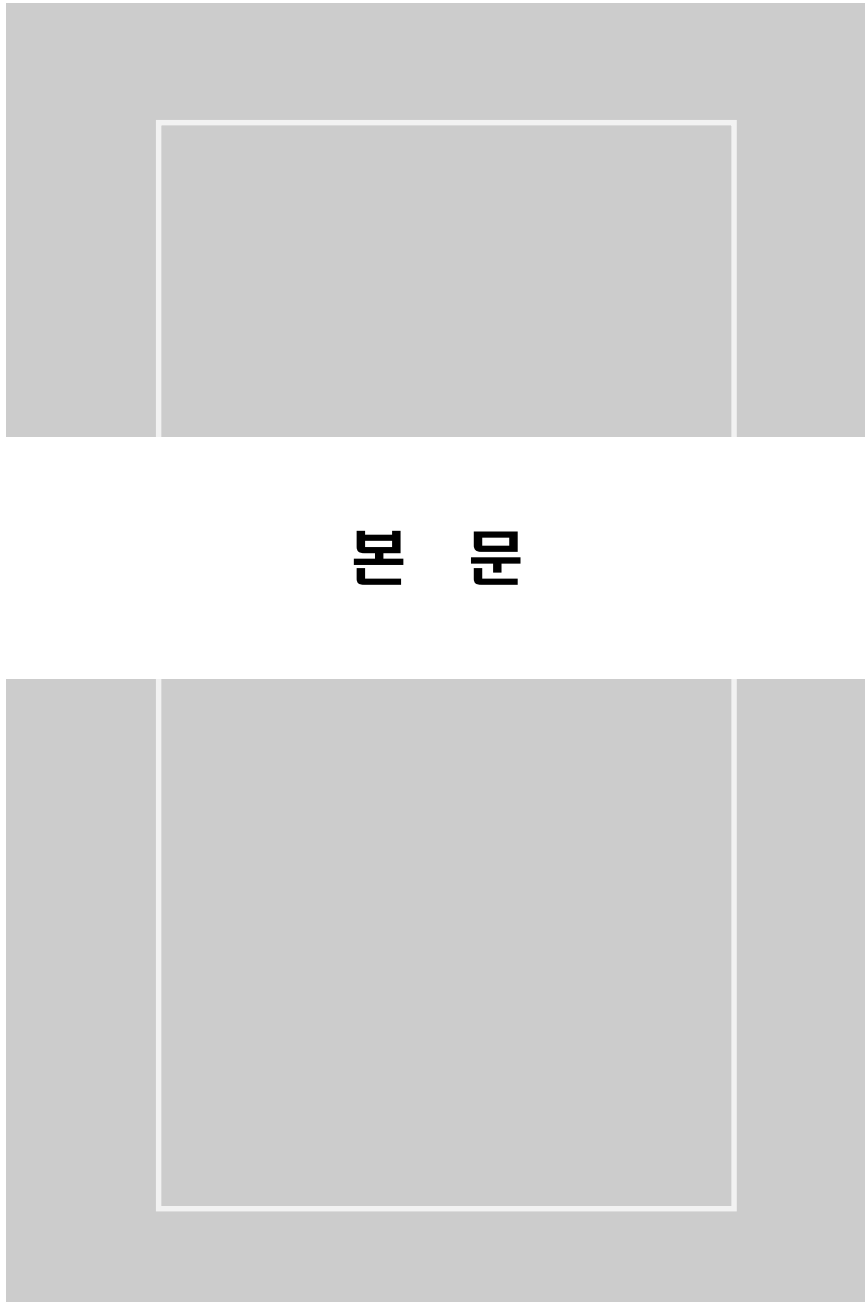
○ **(데이터 신뢰성 확보) 데이터 수집-가공-개방** 각 과정의 **적법성, 투명성 점검** 및 **데이터 이력 관리 방안** 등 도입 추진('22~)

4 데이터 구축에 함께한 분들의 성장을 지원합니다

○ **(단계별 교육 활성화)** 일반라벨러-전문라벨러-관리자 등 **수준별 맞춤형 교육**을 **1만여명**에 제공(~21.12)하고, **전문가·관리자 과정** 점진 **확대**

○ **(타 교육과정 연계)** 데이터 **구축·활용 저변**을 **확대**하고, 데이터 **구축 경험**을 관련 분야 **경력**으로 **발전**시키도록 **학교·직업교육** 등 연계('22~)

○ **(크라우드소싱 적용 분야 확대) 단순 데이터 수집, 가공 작업**에서 **데이터 검수·품질관리** 등 **전문성**이 필요한 분야로 **확대** 추진('23~)



순 서

Ⅰ. 고도화 추진배경

◇ 디지털 전환 가속화에 선제적 대비 → 데이터 댐 구축 착수

□ 정부는 **'인공지능 국가전략'**('19.12), **'디지털 뉴딜'**('20.7)을 추진하며 **디지털 경제**로의 **전환 가속**을 위한 **정책 노력 본격화**

※ '국가 AI R&D 전략계획(美, '19)', '디지털전략 정책안(EU, '20), '국가 인공지능 전략(英, '21) 등 주요국도 인공지능·데이터 기반 디지털 경쟁력 증진을 위한 국가적 대응 강화 중

□ 특히, 디지털 뉴딜 대표 프로젝트로 **'데이터 댐' 구축**에 착수('20.9)하여, 디지털 전환의 핵심자원인 **데이터 확보**에 **정책역량 집중**

◇ 인공지능 학습용 데이터 구축을 중점 과제로 투자 대폭 확대

□ **데이터 댐 주요사업**으로 인공지능 개발·고도화에 필수적인 **인공지능 학습용 데이터 구축**에 **'25년**까지 **약 2.5조원**의 **대규모 투자** 추진

□ 이를 통해, 각 분야의 **인공지능 도입 확산**과 **기술 발전**을 선도할 **데이터 수요**를 **발굴**하고 면밀한 검토를 거쳐, **대규모 추가 구축·개방***

* 기 구축된('17~'19) 21종에 더하여 '20년 구축한 170종의 데이터 추가 개방('21.6)

◇ 투자 효과 극대화를 위한 고도화된 데이터 확보 전략 추진 필요

□ **데이터 댐 착수 1년**을 맞이하여, **정책 효과 극대화**를 위해 그간의 경험을 토대로 **체계적인 인공지능 학습용 데이터 확보방안** 마련 필요

□ 아울러, **구축한 데이터**의 **활용 편의**와 **가치**를 더하는 한편, 데이터 구축 **참여인력**의 **전문성 증진**을 위한 지원 강화도 요구되는 시점

➡ 그간의 **경험**을 **토대**로, 앞으로의 **기술·산업 발전상**과 **연계**하여 **데이터 구축·개방, 활용** 전반에 대한 **전략적 방향 제시**가 긴요

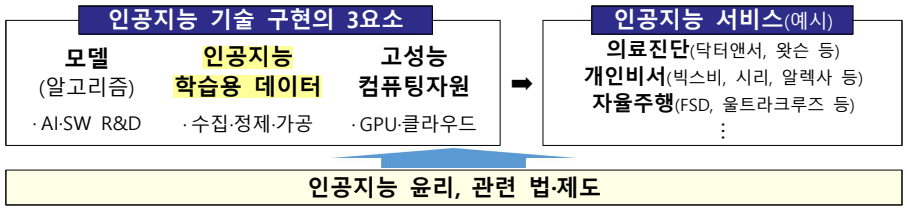
Ⅱ. 인공지능 학습용 데이터 구축의 필요성

◇ 인공지능은 데이터를 포함한 3가지 핵심 요소를 통해 구현

□ **인공지능**은 **인간**의 **지능**을 **컴퓨터**로 **구현**하는 기술로 ①**상황 인지**, ②**이성·논리적 판단**과 **행동**, ③**감성·창의적 기능**을 수행하는 능력

※ 시청각·인식·동작 등을 구현하는 **기본지능**을 기반으로 챗봇상담·질병진단·재난예측 등 다양한 분야의 **응용지능**을 개발하고, 창작·개발 등이 가능한 **창의지능**으로 고도화

□ 인공지능은 명령어 집합인 **모델**(알고리즘), 모델이 성능확보·강화를 위해 학습하는 **데이터**, 대용량 연산에 필수적인 **컴퓨팅자원**을 통해 구현



◇ 인공지능 학습용 데이터는 분야별 인공지능 확산·발전의 토대

□ **인공지능 학습용 데이터**는 **모델**이 정확하게 **인식·학습** 가능한 **텍스트·영상·이미지·음성** 등의 **원천데이터**와 **가공**(라벨링)**된 데이터**를 총칭

< 인공지능 학습을 위한 데이터 가공 예시 >

음성·텍스트(한국인 대화음성)	**이미지**(질환 이미지)	**영상**(CCTV 이상행동분석)
		
음성파일을 텍스트로 변환하여 성별, 발화주제 등 속성값 부여	병변 진단·판독, 환부 영역표시·주석 등 정보 기입	영상 속 객체 구분과 이상행동 여부 파악, 동작 등 설명·표기

□ **대량의 데이터 학습**은 **모델 성능향상**으로 **직결***되나 **확보·가공**에 많은 **시간·비용 소요**** → **다양한 양질의 데이터 확보**가 중요한 **선결과제**

* (예) 학습용 데이터 이미지가 많을수록 객체분류 인공지능의 정확도 향상('20, 구글)
: 1.26백만개 이미지(94%) → 1.4백만개 이미지(94.5%) → 3백만개 이미지(97%)

** 국내 AI·데이터 기업은 AI 개발시간의 80%, 비용의 75%가 데이터 확보에 소요된다고 응답('20, NIA)

◇ 미국 등 선도국은 기업·플랫폼 중심으로 데이터 구축·활용 활성화

□ 구글·AWS·MS 등 **빅테크 기업**을 중심으로 **인공지능 학습용 데이터 구축·활용**이 활성화되고 있으며, **데이터 가공 시장** 등 **성장**을 견인

○ **데이터셋 가공 플랫폼**을 제공하는 **기업들**이 활발히 **성장 중**으로, 일부 인공지능 기업은 **해당 기업 인수·합병**을 통해 산업 성장에 대비

※ (Scale AI Inc.(美)) 로봇·자율차·드론용 이미지에 주석을 다는 SW 개발, 1,800만불 펀딩 유치
(Appen(濠)) 크라우드소싱으로 학습용 데이터 생산·제공, 美 AI데이터 기업 인수에 3,500억원 투자

□ Kaggle, 이미지넷 등 민간의 **인공지능 개발자 경진대회 플랫폼** 등을 통한 **데이터 축적·공유·개방** 등도 활성화

◇ 우리도 관련 기업·시장이 태동 중이나 데이터 부족 문제 지속 제기

□ 우리도 **인공지능 학습용 데이터 확보**를 위한 **정부 투자 확대** 등으로 **데이터 구축·가공** 관련 **기업, 시장** 등이 **탄생·성장** 중

○ 이를 통해 **국내 인공지능 서비스 개발**에 적합한 **다양한 데이터 자원**이 **점차 확충**되고, **기업**의 **데이터 구축 수요**도 발생

□ 다만, 아직 많은 기업들이 **인공지능 도입·운영**에 대한 **가장 큰 어려움**으로 **데이터 부족 문제**를 지적*하고 있어, **지속 개선 필요**

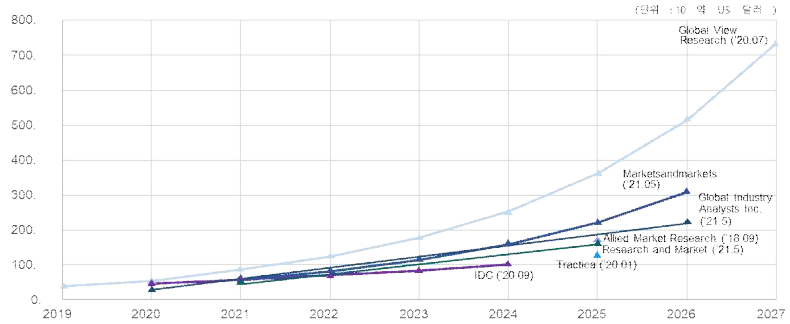
* 인공지능 도입·운용 시 어려운 점('20, KDI) : 데이터 부족(36.1%) > 예산·인력 부족(30.6%) 등

➡ ❶**국내 실정**에 맞는 **인공지능 학습용 데이터**의 조속한 **양적·질적 확충**,
❷**데이터 확보·가공 역량이 부족**한 인공지능 관련 **중소·스타트업, 학계, 개발자** 등을 위한 **정부 주도**의 **데이터 구축·개방 확대** 요구 지속

참고 국내외 인공지능 시장 전망

□ 글로벌 인공지능 시장 전망

○ 인공지능 시장은 **기술 발전, 서비스 상용화·고도화**에 따라 **성장**을 **가속화**하여, **'25년**까지 **208~374조원 규모**로 성장할 것으로 전망

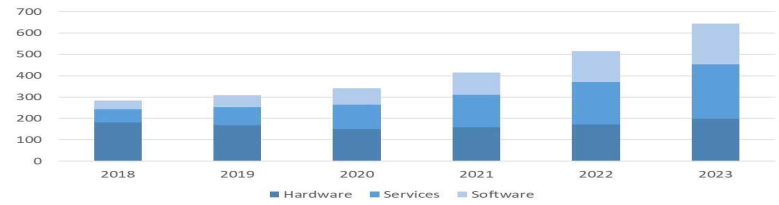


○ **인공지능 SW**가 **전체 시장**의 **88%**에 달하며('21, 한국IDC), **인공지능 플랫폼 시장**이 **연평균 45% 이상**의 **고도성장** 예측('21, Tractica)

□ 국내 인공지능 시장 전망

○ 국내 인공지능 시장은 '19년 약 1.5조원 → **'25년 10.5조원 규모**로 **빠르게 성장**(연평균 38.4%)할 것으로 예상('20, 한국신용정보원)

< AI 제품 유형 별 시장 규모(단위: 10억원) >



○ **머신러닝·컴퓨터비전·자연어처리** 등의 기술이 **성장을 주도***하고, **'22년 이후** HW보다 **서비스·SW 시장**이 **더 커질 것**으로 예측**

* '25년 머신러닝(4.24조원) > 컴퓨터비전(2.64조원) > 자연어처리(2.49조원) 순('20, 한국신용정보원)

** '18~'22년(5년간) 국내 인공지능 서비스·SW시장 연평균 30%이상 성장 전망('20, 한국IDC)

Ⅲ. 현황 및 추진방향

◇ [현황] 디지털 뉴딜 : 인공지능 · 데이터 생태계 활성화 기폭제 마련

□ **디지털 전환 가속화**에 **대비**하고 코로나19로 인한 **경제위기**를 **극복**하고자 '**데이터 댐**' 구축*을 중심으로 **디지털 뉴딜** 본격 추진('20.7~)

* **(데이터 댐 7대 사업)** ❶인공지능 학습용 데이터 구축, ❷빅데이터 플랫폼·센터 구축, ❸데이터 바우처 제공, ❹인공지능 바우처 제공, ❺인공지능 융합 프로젝트(AI+X), ❻클라우드 플래그십 프로젝트, ❼클라우드 이용 바우처 제공

○ **각 분야**의 **인공지능 도입·활용**을 **조기에 확산**할 수 있도록 **'25년**까지 **1,300종**에 달하는 **인공지능 학습용 데이터 구축·개방** 추진

연 도	2020	2021	2022	...	2025
종수(누적)	191	381	691	...	1,300

□ 디지털 뉴딜 1년여간 **대량의 인공지능 학습용 데이터 구축·개방, 활용 확산, 품질관리체계 마련·적용** 등 소기의 성과 달성

❶ **(대규모 데이터 추가개방) 수요분석, 전문가 검토**를 거쳐 **구축**한 음성·자연어, 비전, 헬스케어 등 **8개 분야 170종 4.8억건 개방**('21.6)

※ '20년 데이터 구축 과정에 기업·대학·병원 등 571개 기관과 청년, 경력단절여성, 장애인 등 4만여 명의 인력이 참여

❷ **(활용과 서비스 창출 확대)** 총 11만여 건의 다운로드 중 **170종 추가개방**('21.6) 후 **6개월간 5.8만건**에 달하며, 다양한 **혁신서비스 창출**에 기여

라이드플럭스	· **자율주행 데이터** 기반 **무인이동 서비스 도입·고도화** 및 **지역 확대** ※ 제주관광지 자율주행 셔틀('21.8), 세종정부청사 승객이동 서비스('21.9) 운영
아워랩	· **수면 질 분석 데이터**를 활용한 **수면자세 감응형 수면무호흡 치료기기 개발** ※ 식품의약품안전처 의료기기 품목허가 취득('21.2)
씨유박스	· **개인동의 기반 안면이미지 데이터**를 활용한 **위변조 감지 알고리즘** 개발 ※ 국내 기업 최초 관련 기술 국제표준(ISO/IEC 30107-3) 인증 획득('21.7)

❸ **(데이터 품질관리체계 정비)** 데이터 **품질 강화**를 위해 **데이터 구축 전 과정**에 **참조·활용**할 수 있는 **각종 품질기준*** 마련·적용('20.10~)

* 품질공통기준(수집·가공·검수 등에 필요한 요구사항), 참조기준(데이터 속성 정보, 라벨링 방법 등)

참고 인공지능 학습용 데이터 구축사업 개요

□ 사업개요

○ (개요) **'디지털 뉴딜'** 핵심프로젝트인 **'데이터 댐'**의 **대표사업**으로 **인공지능 서비스 개발**에 필수적인 **학습용 데이터**를 **대규모**로 **구축·개방**하는 사업

○ (목적) **양질의 인공지능 학습용 데이터**를 **대규모**로 **구축**하여 중소·스타트업 등 **민간**의 **인공지능 기술개발 촉진** 및 **관련 산업**을 **육성**하고,

- **일자리 창출** 등 **민간 참여 기반**의 **인공지능·데이터 선순환 생태계** 조성

○ (사업내용) 인공지능 학습용 데이터를 구축·개방하고자 하는 **기업**, **대학**, **출연연**, **공공기관**, **협회**, **지자체** 등 **민간·공공 법인** 등을 지원

※ (디지털뉴딜 목표 사업기간) ~'25년 / (디지털뉴딜 목표 사업비) 약 2조 5천억원

□ 추진경과

○ (구축·개방) **'20년 170종 추가 구축**을 완료('21.2)하고 **품질검증**, **개발자·전문가 대상 사전공개**를 통한 **오류·유효성 검수** 후 AI Hub(aihub.or.kr)에 **개방**('21.6)

- **'19년**까지 법률, 특허, 한국어 음성, 이상행동 CCTV 등 **21종 4,650만건**, **'20년 170종 4억 8,000만건**의 데이터를 추가 구축·개방

< 인공지능 학습용 데이터 구축·개방 현황 >

○ (품질관리) **품질 공통기준** 마련·적용('20.9~), 8개 분야 80여명의 **품질자문단** 운영('20.10~)

○ (일자리 창출) **직접고용**과 함께 국민 누구나 참여할 수 있는 **크라우드소싱** 방식 적용을 통해 **총 40,165명**('20년 기준)의 일자리 창출

◇ [추진방향] 풍부한 수요를 바탕으로 분야별 전략적 데이터 확충

1. 충분한 과제 수요확보와 종합적 검토가 가능한 체계로 운영 필요

평 가	**매 사업 추진 시 개별 수요**를 기반으로 **과제 선정**을 추진하여, **다양한 시기별 데이터 수요**에 대한 **포괄적 검토**에 **한계**
추진 방향	▪ 신규로 파악된 수요 외에 **기존 미선정 과제**도 **재검토**할 수 있도록 하여, **제기된 모든 수요**를 **'과제 수요 풀'**로 **활용** ▪ **기존 구축 데이터**의 **확장·보완 수요** 등도 **종합적**으로 **고려**한 **연도별 구축과제 선정** ☞ '20년 추경 이후 제기된 **3,328건**에 더하여, 데이터 수요 풀 지속 확대 : 신규수요 + 이전 사업 미선정 과제 + 기 구축 데이터 확장·보완수요 등

2. 디지털 뉴딜에 따른 1,300여종에 대한 전략적 구축방향 설정 필요

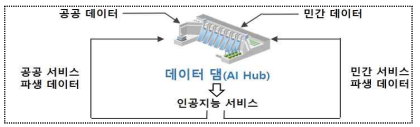
평 가	**각 사업 기간별 분절적 추진**으로, **全산업·사회적 디지털 전환**을 **촉진**할 데이터 구축·활용에 대한 **체계적 방향성 제시**가 **미흡**
추진 방향	▪ **디지털 뉴딜**(~'25)로 추진되는 **학습용 데이터 구축 사업 목표**를 **'민간 주도**의 **데이터 생태계 형성'** 지원으로 명확화 ▪ 이를 위해, 매년 확장되는 수요기반 데이터 풀을 활용하여 **1,300여종**에 이르는 데이터를 **각 분야**에서 **전략적·효율적**으로 **구축·활용**할 수 있는 **로드맵 제시**

3. '26년 이후의 데이터 생태계는 자생적 선순환 모델을 지향

□ 기 구축 또는 구축 중인 **학습용 데이터**의 **활용성 증진**과 함께, **데이터 댐**이 **지속 성장·고도화**할 수 있는 **선순환 모델** 지향

○ 디지털 뉴딜 후에도 **전략적 확충**이 필요한 **데이터 확보**, **수요자 편의**와 **데이터 활용성 증진**을 위한 **지원 대책**은 **지속 추진** 필요

□ **공공·민간** 부문별 **구축 데이터**, **인공지능 서비스**에서 **파생**되는 **데이터**가 **데이터 댐**으로 **환류**되는 체계로 운영



참고 인공지능 학습용 데이터 수요기반 과제 추진

□ **디지털 뉴딜** 추진 후, **4차례**의 **수요조사**를 통해 **총 3,328건**의 **데이터 수요**가 **축적**되었으며, **매년 과제기획위 검토**를 거쳐 **우선 추진과제**를 **선별**

연도		조사 기간	수요	구축량	조사대상
'20년	추경사업	'20.4~6월	1,251	150	부처, 지자체, 공공기관, 연구기관, 데이터·AI 기업 등
'21년	본사업	'20.10~11월	891	150	
	추경사업	'21.3~4월	172	40	
'22년	사업	'21.9~10월	1,014	310	
계			**3,328**	**650**	

○ '20년까지 구축된 **191종**의 **수요기반 데이터**를 **구축·개방 중**이며, **'21년 190종**의 신규 데이터 **추가 구축** 및 '22년 대비 **추가 수요조사·검토** 추진 중

< 인공지능 학습용 데이터 수요와 선정 과제 예시 >

: 선정 과제 / : 수요제기(미선정) 과제

한국어	영상·이미지	헬스케어	안 전
다화자 음성합성 데이터 소음환경 음성인식 데이터 주제별 일상 대화 데이터 한·영 번역 병렬 말뭉치 ⋮	일상생활 영상 데이터 마스크 착용 안면 이미지 데이터 한국인 형상치수 측정 데이터 CCTV 유동인구 데이터 ⋮	심초음파·심전도 데이터 캡슐내시경 이미지 데이터 의료분야 음성 데이터 임산중 태아 초음파영상 데이터 ⋮	보행안전 도로 시설물 데이터 산림 수종 이미지 데이터 기상 정보 데이터 해안 오염물질 데이터 ⋮
미디어 콘텐츠 텍스트 데이터 교통안내 다국어 말뭉치 금융 용어 말뭉치 ⋮	모의면접 영상 데이터 지하철 부정 승하차 데이터 실내 가구 배치 데이터 ⋮	식습관 등 관련 혈당 데이터 X-ray 이미지(관절질환) 환자 생체 신호 데이터 ⋮	시설물 손상 영상 데이터 자살시도자 이동 패턴 데이터 네트워크 공격 트래픽 데이터 ⋮

농·축·수산	자율주행	제조·관광 등 주요 산업
정밀농업 노지작물 통합데이터 스마트 양식장 통합데이터 동의보감 약초 이미지 데이터 한우 산체중실시수 등급 데이터 ⋮	교통사고 영상 데이터 멀티센서 동선 추적 데이터 탑승자 상황 인식 영상 데이터 지자체 도로 정비 데이터 ⋮	**(제조)** 주요 산업군 공정별 부품 이미지 산업현장 고숙련자 현장작업지식 데이터 등
과일 및 작물 이미지 데이터 양식장 공간 정보 데이터 ⋮	도로별 통행량 시계열 데이터 온라인 정기배송 데이터 ⋮	**(관광)** 문화재 현판·비문 등의 한자 데이터 관광 관련 챗봇 시나리오 데이터 등
		(교육) 손글씨 수학문제 및 풀이 이미지 데이터 각종 자격증 기출문제 데이터 등
		(문화) K-POP 안무 영상 데이터 한국인 얼굴 구성 요소별 데이터 등
		⋮

Ⅳ. 추진전략

비전 풍부한 데이터 자원 기반, 인공지능 기술·산업 선도국 도약

성과 목표 (~'25)

	2021	2022	2023	2024	2025
데이터 구축	누적 340종		900종	1,100종	1,300종+α
데이터 다운로드(활용)	누적 25만건		70만건		100만건
인공지능 개발·확산	누적 3천건		8천건		1만건

추진 과제

1 데이터 구축·개방 로드맵 수립·이행

❶ **체계적 데이터 확보를 위한 방향과 전략을 제시합니다.**
인공지능 발전·확산 방향 + 분야별 특성을 고려한 전략분야 설정

❷ **분야별 데이터 구축·개방, 이렇게 추진됩니다.**
기반기술(여러 기술·산업에 공통되는 인공지능 구현지원)
+ 전략분야1(공익 증진과 조기 성과창출) + 전략분야2(산업·영역별 혁신) 중심의 확장

- 기반기술: 한국어, 영상·이미지
- 전략분야1: 헬스케어, 교통·물류, 재난·안전·환경, 농·축·수산
- 전략분야2: 제조, 에너지, 교육, 금융, 국방, 법률, …(+α)

❸ **로드맵 이행체계를 갖추어 나갑니다.**
분야별 전문성을 갖춘 각 부처 참여·협력 확대

2 구축 데이터의 활용성과 가치 증진 지원

❶ 쉽게 접근하고, 편리하게 활용할 수 있도록 합니다.
❷ 다양한 정부 지원사업과 연계하여 활용가치를 높여 갑니다.
❸ 믿고 쓸 수 있는 양질의 데이터로 만들어 갑니다.
❹ 데이터 구축에 함께한 분들의 성장을 지원합니다.

1 데이터 구축 · 개방 로드맵 수립 · 이행

1 체계적 데이터 확보를 위한 방향과 전략을 제시합니다

□ 기반기술 + 2대 전략분야를 축으로 종적 · 횡적 확장

① (기반기술) **한국어, 영상·이미지** 등으로 **여러 기술·산업**의 **공통 기술**로 활용되는 **범용적 성격**의 **인공지능*** **구현**을 지원하는 분야

* 시각·청각·인식·동작 등을 비롯하여 다양한 분야에서 활용될 수 있는 공통 기술

② (전략분야1) 헬스케어·교통·안전 등 **공익 증진**에 기여할 수 있고, **조기에 인공지능 확산**과 **성과 창출**을 **기대**할 수 있는 분야

③ (전략분야2) 인공지능을 통한 **산업·영역별 혁신**을 **촉진**할 수 있는 분야

< 로드맵 수립 분야와 분야별 세부 영역 >

기반기술	한국어	**음성**(음성인식 · 합성), **자연어**(대화문, 자연어 요약 · 생성, 질의응답, 번역), **기반 데이터**(자연어 기반기술, 지식베이스)			
	영상·이미지	**대상인식, 영상이해, 동작·영상의 지능적 인지, 3D 비전, 멀티모달** (※복합적인 정보가 포함된 데이터)			
전략분야1	헬스케어	**질병**(중증질환(암), 만성질환, 장기별질환, 특수질환), **헬스케어**(범용 대규모 의료데이터, 건강인 헬스케어), **사회적 의료문제, 첨단의료**			
	교통·물류	**자율주행, 지능형 신호체계, 스마트 물류**			
	재난·안전·환경	**재난**(자연재난·사회재난), **안전, 환경**			
	농·축·수산	**농업**(스마트팜, 디지털육종, 미래 소재·기술), **축산, 수산**			
전략분야2	제조	로보틱스	법률	교육	+ α
	금융	지식재산	문화·관광	스포츠	

□ 로드맵 수립 시 핵심 고려사항

① 각 **분야별 인공지능·데이터** 관련 **주요 정책, 기술로드맵** 상 계획에 따른 **기술 수준**과 **발전 방향**을 고려한 **데이터 확보 전략**으로 수립

② 전략분야·종별 **시급성, 중요성, 파급효과**와 **인공지능 기술 트렌드**를 함께 고려

③ **'25년**까지 **디지털 뉴딜**에 따른 **1,300여종**의 데이터를 구축하고, **매년** 변화하는 **정책환경**을 **반영**하여 **차년도**의 **구체적 추진과제 기획·선정** 추진

< 고도화 방안 이행을 통한 데이터 댐 기반 선순환 생태계 활성화 >

참고 인공지능 학습용 데이터 구축 로드맵 수립 추진경과

1 인공지능 학습용 데이터 구축 로드맵 핵심 분야 발굴 ('20.10~11월)

○ 인공지능 관련 **정책·시장 동향 분석***을 바탕으로 **후보 분야**를 **선정**(20개)하고, **산·학·연 전문가**(50여명) 대상 **가치평가****를 통해 **주요분야** 도출

* 국내외 인공지능 전략, 관계부처별 AI 관련 기술개발·산업진흥 계획, 사업 등
** (평가기준) AI 활용 적합성, AI 학습데이터 구축 필요성, 정부투자 타당성 등

- **핵심기반**(한국어, 영상·이미지), **전략분야**(헬스케어, 교통·물류, 농·축·수산, 재난·안전·환경) 外 분야(제조, 교육, 문화 등)는 **수요**에 **기반**하여 과제를 **지속 발굴**할 **방침**

2 핵심 분야별 구축 로드맵 초안 도출 및 과제 추진 ('20.12~'21.5월)

○ 핵심 분야별로 **산·학·연 전문가, 각 부처 추천위원**이 **참여**하는 **분과위원회**를 **구성·운영**(총괄기획위, 분야별 분과위원회, 80여명)하여 **분야별 로드맵 초안 도출**

- 각 **부처·지자체** 및 **전문기관**, 인공지능 분야 **협·단체, 활용기업** 및 **개발자 커뮤니티** 등의 광범위한 **수요조사**('20.10~11월) **결과**를 **검토·반영**

※ 협·단체(231명), 부처·전문기관(24개), 인공지능 개발자 커뮤니티(345명) 등으로부터 891개(민간 751개, 공공 140개) 신규과제 수요 제기

○ 도출된 **로드맵 초안**에 대한 **추가 의견수렴**(온라인, '21.12~1월) 및 **총괄기획위원회**를 거쳐 **핵심분야**와 **세부 영역** 등을 선정

- **로드맵 초안**을 바탕으로 '21년 **총 190종**(본예산 150종, 추경 40종)의 **우선추진 과제**를 기획하여 **데이터 구축 중**(본예산 '21.3월~, 추경 5월~)

3 의견수렴 및 로드맵(안) 마련 ('21.6~11월)

○ **로드맵 초안**에 대해 **관계부처** 및 구축·활용 **전문기업** 의견 수렴, **산·학·연 전문가 검토**를 거쳐 **인공지능 학습용 데이터 구축 로드맵(안)** 마련

○ 수립된 분야별 **로드맵(안)**을 토대로 **'22년 310종**(잠정)의 데이터 등에 대한 **신규 구축** 기획·추진

- 로드맵 **세부추진방향**은 **정책환경 변화**를 고려하여 **지속 보완** 예정

2 시장이 원하고 국가가 필요로 하는 데이터, 이렇게 구축됩니다

기반기술 **한국어**

"다양한 분야와 상황에 활용할 수 있는 한국어 인공지능 개발을 뒷받침합니다"

□ 데이터 개요와 필요성

○ **(개요) 한국어 음성·문자**를 **컴퓨터**가 **이해·처리·구사**할 수 있도록 **가공**된 데이터로 **음성인식, 요약·합성, 번역, 질의응답** 시스템 등에 활용

※ (활용서비스 사례) 다양한 업종별 대화형 챗봇, 전문영역 등의 회의록 요약과 화자 구분, 방언·저음질 음성 등을 인식하는 AI 스피커, 주요언어-한국어 간 통·번역서비스 개발 등

○ **(필요성) 한국어 기반**의 **모든 인공지능 구현**에 필요한 데이터로 **상황·용도·개인별 편차**가 크기 때문에 **다양한 유형·수준별 대규모 데이터 축적**이 요구되며, 각 국도 언어 데이터 확보에 역량 집중

※ 초거대 인공지능 기반의 서비스 개발 및 성능 향상을 위한 기능별 한국어 데이터 필요

□ 주요 세부과제별 목표

	As-Is	To-Be	
음 성	단일한 분야·업종·환경의 음성 데이터 구축	소음·저음질·다화자 등 다양한 환경을 반영할 수 있고, 복합적인 정보를 담은 음성 데이터로 확장	각종 음성정보 인식·해석 능력 증진
자연어	간단한 대화문, 질의응답, 번역문 말뭉치 등 유형별 텍스트·이미지 등 확보	다채로운 분야의 목적·상황별 언어를 인지하고, 요약·합성 기능을 구현·고도화할 수 있는 데이터 구축	언어 분석, 요약·생성능력 고도화
기 반 데이터	전문분야 언어의 의도·의미 분석을 지원하는 데이터 구축 ※ 법률·특허 용어 분석·라벨링 데이터 등	각 산업·사회 분야별로 높은 수준의 언어 인식·해석을 지원할 기반데이터 및 도구 확보	다양한 분야의 한국어 의미분석 기능 강화

☞ '모두의 말뭉치'를 운영 중인 문화체육관광부(국립국어원)와 전략적 협력체계 구성·운영

□ 한국어 데이터 로드맵

	2021	2022	2023	2024	2025
	선진국 수준의 언어 이해 기술 확보		분야별·산업별 음성 인식과 합성이 가능한 인공지능 개발		
음 성	특정 환경의 음성 인식 데이터	다양한 개별 업종·분야, 환경의 음성 인식·합성 데이터		멀티모달 음성 인식·합성 데이터	
자연어	다양한 분야·목적·상황별 대화, 영상·이미지 해석, 대화·회의록 기반 문서 요약·생성 데이터			말뭉치 생성 데이터, 각종 전문지식·SNS·업무 데이터	
기 반 데이터	분야·업종별 원시데이터* * 컴퓨터가 판독·해석할 수 있는 기초데이터		분야·업종별 지식그래프	다양한 용어정보와 원시데이터·지식그래프 간 통합 데이터	

☞ '21년까지 한국어대화·방언·논문·도서요약·콜센터질답·번역말뭉치·법률·특허 등 49종 구축·개방

기반기술 영상 · 이미지

"폭넓은 수준·종류별 영상·이미지 제공으로 인공지능의 시야를 한층 넓혀줍니다"

□ 데이터 개요와 필요성

○ **(개요) 컴퓨터**가 **사물·상황** 등의 **의미**와 **정보**를 **감지·인식·분석**할 수 있도록 지원하는 **다양한 종류·유형**의 **영상·이미지** 데이터

※ (활용서비스 사례) 인공지능 로봇·기계 등의 동작 인식·구현, 불량품 판독, 3차원 공간 인지·구성 등

○ **(필요성)** 자율주행, AI진단, 스마트공장, 보안시스템 등 **각 산업·사회** 분야의 **시각 기반 인공지능 구현·고도화**에 필수불가결한 자원

□ 주요 세부과제별 목표

	As-Is	To-Be	
대상인식	한국인 안면인식(개인동의기반) 등 기초데이터 확보	추상화된 대상에 대한 인식·분석, 국내 도심 공간정보 등으로 확대	국내 객체·공간 인식능력 향상
영상이해	다양한 형태, 행동 관련 영상 데이터 축적	복합적인 상황, 다양한 객체·배경의 인식·처리를 지원하는 데이터로 확장	다양한 객체의 통합이해능력 제고
동작·영상 지능적 인지	센서로 간단한 동작·인식 등을 지원하는 데이터 구축	동작·상황 등의 종합적 인지, 가상의 공간 인식을 지원하는 데이터 구축	동작인지 및 상호작용능력 제고
3D비전	실내외 공간등의 3차원 모델 데이터 구축	도시 단위, 크고 복합적인 실내환경 등에 대한 종합적 3차원 데이터 구축	파노라마 3D 구현능력 확보
복합상황	다양한 센서, 상황으로부터 수집된 시각적 데이터 확보	대규모, 복합 상황에 대한 시각적 인식 지원을 위한 데이터 확보	복합상황에 대한 이해력 제고

□ 영상 · 이미지 데이터 로드맵

	2021	2022	2023	2024	2025
	현재 적용 가능한 영상·이미지 데이터의 고도화		동영상 및 3차원 세계를 전체적으로 이해·재구성 할 수 있는 인공지능 구현		
대상인식	HD급 해상도의 한국인 안면 데이터, 안면인식 기반 출입 지원 데이터		추상화 인식 지원 데이터	한국 도심 공간정보 데이터	
영상인식	다양한 형태, 행동 관련 데이터	복합적인 상황, 신호·동작에 대한 인식·처리 관련 데이터		다양한 객체·배경·환경에 대한 통합 인식·처리를 지원하는 데이터	
동작·영상 지능적 인지	동기화된 다양한 센서로부터 수집된 데이터, 센서 기반의 주행·동작 인지 관련 데이터			가상·실제 복합상황의 정보·텍스트, 임의상황·다국어 환경 데이터	
3D비전	실내외 공간 2D·3D 영상 데이터	총체적인 3D 재구조화를 지원하는 데이터		파노라마 수준의 3D 재구조화를 지원하는 데이터	
복합상황	다양한 목적의 센서에서 수집된 데이터, 상황에 대한 시각적 데이터			대규모 복합상황 데이터	

☞ '21년까지 한국인 감정·인식 복합영상, 장면·인물 인식 데이터 등 총 38종 구축·개방

전략분야 1-1 헬스케어

"인공지능이 한차원 높은 의료서비스와 건강관리를 실현할 수 있도록 도와줍니다"

□ 데이터 개요와 필요성

○ **(개요)** 한국인이 취약한 **질병 판독·진단**과 일상 속의 **건강 관리**, 소아·노인 등 **취약계층**의 **건강 유지** 등을 뒷받침할 기반 데이터

※ (활용서비스 사례) 질병 판독·예측 정확성·속도 향상, 건강·수면 관리 보조, 가상환경 의료교육 등

○ **(필요성)** 각종 인공지능 기반 **의료기기 개발**과 **서비스 혁신**을 통한 **보건의료** 분야 **사회·경제적 비용 감소, 헬스케어 산업 활성화**에 기여

□ 주요 과제분야별 목표

	As-Is	To-Be	
질 병	시급성 높은 질병 판독, 진단 지원 데이터 구축	임상결정 지원부터 합병증 예측, 재발·악화 예측 등을 위한 데이터 확보	질병 판독·진단·예측 능력 제고
건강관리	건강검진·문진 정보와 수면관리 관련 정보 구축	응급실·중환자·만성질환자 정보와 일반인 식이·스트레스 정보 등으로 확장	건강인 일상 건강관리 기능 향상
사회적 의료	임산부 관리 등 기초적 사회의료 데이터 확보 착수	병원 안전관리와 서비스 개선, 소아·청소년 ·소외계층 의료 관리로 확대	의료 소외계층 서비스 개선
첨단의료	신약개발 지원 등을 위한 기반 데이터 확보 착수	다양한 신약 ·첨단의료기기 개발 지원, 가상환경 기반 의료교육 데이터 등 확보	신약·의료기기 및 의료 교육 혁신

☞ 보건복지부 '한국형 중환자 AI 데이터셋', '암 특화 AI 데이터셋' 등과 중복 방지 및 협업·연계 강화

□ 헬스케어 데이터 로드맵

	2021	2022	2023	2024	2025
	질병 진단 기기 ·솔루션의 고도화		질병을 예측하고 건강증진 자가 관리를 지원할 수 있는 인공지능 서비스 개발		
질 병	조기검진 및 선별진단 데이터, 질병별 진단 및 판독보조 데이터		임상결정 지원 데이터	합병증 및 재발·악화 예측, 사망사고 예측 지원 데이터	
건강관리	건강검진, 의료진·환자 문진, 수면관리 관련 데이터		응급실 및 중환자 관리 데이터, 식이·라이프로그 및 스트레스 평가 데이터	만성질환자 코호트 연구, 치매·노화 관리 데이터	
사회적 의료	임산부 관리, 병원 안전관리 데이터		요양병원의 간호 서비스 관련 데이터, 소아·청소년 우울증 및 소외계층 의료 관련 데이터		
첨단의료	-	유전체 기반 신약개발 데이터	첨단 의료기기 로봇 개발 지원 데이터, VR·AR 기반 의료진 교육용 데이터		

☞ '21년까지 각종 암·질환진단 데이터, 의료진·환자 음성, 인체자세·영상 등 총 35종 구축·개방

전략분야 1-2 교통 · 물류

"작동이 불필요한 수준의 자율주행, 스마트물류 구현을 위한 밑거름이 됩니다"

□ 데이터 개요와 필요성

○ **(개요) 레벨4*** **자율주행** 구현에 필요한 각종 **주행 인지·제어, 지능형 신호체계** 기반 데이터와 **물류 스마트화**에 필요한 **분야·과정별** 데이터

* "고도 자동화" 단계로, 자율주행 작동 구간 내 운전자의 주시 및 작동이 불필요한 수준

※ (활용서비스 사례) 자율주행차량·시스템 구현, 실내 자율주행로봇 개발, 교통상황 인지·안내 등

○ **(필요성) 고도의 자율주행 체계**의 **조속한 구현**을 지원하고, **물류 전 과정**을 **지능화**하여 물류 전반의 **비용·소요시간 최소화**를 유도

□ 주요 과제분야별 목표

	As-Is	To-Be	
자율주행차	도로상의 각종 **객체, 장애물, 주차공간 인식** 등 기반데이터 확보	**악천후**와 각종 **이상 상황, 도로 외 공간**의 각종 **객체 인식**과 **상황판별**을 지원하는 데이터로 확대	**레벨4 수준의 자율주행 구현 지원**
지능형 신호체계	**교통법규 정보, 교통혼잡도**와 **교차로·횡단보도 상황 인지**를 지원하는 데이터 구축 착수	**도로 상태 이상**과 **교통사고 발생 인지**를 지원하는 데이터 구축과 함께 기 구축 데이터 확대·고도화	**국내 도로의 지능형 신호체계 구축 지원**
스마트 물 류	**화물유형, 이동경로** 관련 데이터와 **실내 자율주행 로봇**용 데이터 구축	**물류과정**의 **위험 판단**과 **물류배송 경로관리, 항만관제·택배** 등 **전주기 관리**를 위한 데이터로 확대	**물류 과정 전주기의 스마트화 지원**

☞ 국토교통부 '자율주행 데이터 공유센터', 차량·물류기업별 데이터와 전략적으로 연계 구축

□ 교통 · 물류 데이터 로드맵

	2021	2022	2023	2024	2025
	레벨4 자율주행 구현을 위한 지능형 신호체계 기반 구축		**레벨4 자율주행 구현을 위한 지능형 신호체계 점진적 도입·운영**		
자율주행차	일상객체 인식, 주차공간·장애물 인식 데이터	차량제어를 위한 운전자 반응 데이터, 악천후·이상상황에서의 객체 인식, 운전자 상태인식 등을 지원하는 데이터		도로 외 공간의 객체 인식 데이터, 운전자 심리·정서·특성 데이터	
지능형 신호체계	교통혼잡도 인식 데이터				
	교통법규 위반행위, 교차로 내와 횡단보도의 이상 상황, 도로상태 이상, 교통사고 발생 인지 데이터				
스마트 물 류	화물유형 및 이동경로 데이터, 화물차 주행 및 컨테이너 관리 데이터		물류과정 위험 판단 관련 데이터, 최적 물류배송 경로와 항만시스템 관제 데이터		
	실내 자율주행 로봇용 데이터	실외 자율주행 로봇용 데이터		택배 고객 서비스 고도화 데이터	

☞ '21년까지 주행중인 객체, 차량위치측위, 차선·횡단보도 인지영상, 항만구조물 영상 등 총 23종 구축·개방

전략분야 1-3 재난 · 안전 · 환경

"신속·정확한 재난 예측과 선제적 대응이 가능한 시스템 구현을 지원합니다"

□ 데이터 개요와 필요성

○ **(개요)** 홍수·태풍 등의 **자연재난**, 유해물질 유출 및 수질오염 등 **사회재난, 안전사고 분석·예측** 및 **예방·대응·회복**을 지원할 데이터

※ (활용서비스 사례) 재난·범죄 인식 등이 가능한 지능형 CCTV 개발, 시설물 균열 등 안전성 탐지, 산림·해양 오염과 확산 예측·인식 등

○ **(필요성)** 각종 **재난**과 **사고, 환경문제**에 따른 **피해**를 **최소화**할 수 있는 **지능적 예측·대응·체계**를 마련하기 위한 기반 데이터 확보 시급

□ 주요 과제분야별 목표

	As-Is	To-Be	
재 난	풍수해·기후변화·미세먼지 등 **개별 재난 현상, 각종 기반 시설 피해** 관련 정보 구축	**복합적인 자연·사회 재난**에 대한 **예측·대응** 관련 데이터, **디지털 트윈**과 연계한 **시설안전 데이터** 등 구축	**대형·복합 재난대응기술 확보**
안 전	**취약분야, 고위험 산업 분야 사고·인지** 관련 데이터 구축	**각종 생활안전사고** 및 **주요 산업 안전사고** 관련 데이터와 **범죄 인식·예측** 관련 데이터로 확대	**각종 생활안전 선제적 대응 기술 확보**
환 경	**피복상태, 산림수종, 도시환경** 등 국내 환경 기초정보 확보	**기후변화, 수질·해안오염** 양상 등의 **예측·대응**을 위한 데이터와 **국토자원·환경 전반**에 대한 정보 구축	**국토자원·환경에 대한 지능적 관리**

□ 재난 · 안전 · 환경 데이터 로드맵

	2021	2022	2023	2024	2025
	인공지능 기반 개별 재난 인식기술 고도화		**인공지능 기반 대형 복합재난 대응기술 개발**		
재 난	풍수해, 지진 현황, 피해 관련 데이터	자연재난 인식·예측 관련 데이터, 극단적 기후변화 현황·피해 관련 데이터		복합 자연재난 대응 관련 데이터, 재난대응 로봇 개발 지원 데이터	
	미세먼지·수질 인식, 기반시설 관련 데이터	기반시설 위험, 환경재난 인식·예측 관련 데이터		복합 사회재난 대응 관련 데이터, 디지털 트윈 정보 연계 데이터	
안 전	스마트관제·보안 지원 데이터, 안전 취약계층과 산업별 사고 인식·예측 데이터			각종 생활안전사고 대응 데이터, 일상생활의 범죄 인식·예측 데이터	
환 경	토지피복, 산림수종, 폐기물, 도시환경 데이터	기후변화 인식·예측 데이터, 해안오염·물순환 관련 데이터		도로 외 공간의 객체 인식 데이터, 운전자 심리·정서·특성 데이터	

☞ '21년까지 상하수도·폐기물·수질오염·노후시설, 각종CCTV, 안면인식, 산림피복 등 총 31종 구축·개방

전략분야 1-4 농 · 축 · 수산

"생산-유통 효율 극대화, 종사자 편의 증진을 함께 실현할 수 있도록 해줍니다"

□ 데이터 개요와 필요성

○ **(개요) 생산성 향상, 비용·노동력 절감** 등을 지원할 **스마트농업 데이터, 사육** 및 **질병 진단·처방 정보** 등 **디지털 축·수산 실현**을 지원할 데이터

※ (활용서비스 사례) 각종 작물 생육·질병 관리, 축사 운영 자동화와 최적 관리, 스마트 양식장 구현 등

○ **(필요성)** 인공지능 기반 **농·축·수산업 분야 모델 개발**과 **국내외 확산**을 통해 **주요 농·축·수산물 생산-유통** 전반의 **효율화**를 유도

□ 주요 과제분야별 목표

	As-Is	To-Be	
농 업	정밀농업과 디지털 육종 구현을 위한 생육·유전 관련 기초데이터 확보	생육·환경 조절과 생산량 예측, 기기·장비 자동제어 등을 구현할 기반 데이터 확보	스마트팜 모델 개발과 국내외 확산
축산업	축사환경, 사료관리·수급 등 기초데이터 확보	축종별 맞춤형 축사·질병 관리와 지능형 유통·소비 품질관리를 지원하는 데이터 확보	인공지능 기반 가축관리 모델 개발·확산
수산업	양식장 모니터링·관리를 위한 기초데이터 확보	기후 환경, 양식생물 상태, 종자 품질 진단·예측과 효율적인 먹이·수질 관리 등을 위한 데이터 확보	데이터 중심 수산양식 서비스 개발·확산

☞ 농림축산식품부, 해양수산부 등 전문성을 갖춘 부처가 과제 기획, 운영 등 총괄 추진

□ 농 · 축 · 수산 데이터 로드맵

	2021	2022	2023	2024	2025
	주요 품종 중심의 생육관리 지원체계 구축		인공지능 기반의 생육 예측·관리, 축수산 질병관리를 통한 생산성 향상		
농 업	최적 생육조건 모니터링과 주요작물 생육 진단·예측 데이터, 자율형 스마트팜 통합 제어 데이터, 각종 유전자원·식물·병해·재배환경 등에 관한 데이터			생육환경·에너지소요 등 예측 데이터, 자율주행 농기계 작동제어 데이터, 식품·품질·성분 및 발병예측 데이터	
축산업	축사환경 모니터링 및 사양 데이터, 사료개발 지원 데이터		축종·사료·분뇨처리방식별 CO_2 배출량 데이터, 질병진단·예측관리와 백신·의약품 개발 지원 데이터, 유통·소비 통합 품질관리를 위한 데이터		
수산업	기후변화 모니터링 데이터, 백신·의약품 개발 지원 데이터, 양식장 환경·사양과 양어행동 모니터링 데이터, 종자 입식 및 사료 공급, 종자 유전체 데이터			유통·소비 통합 품질관리를 위한 데이터 어병 진단예측 데이터 양식장 환경 및 사료별 CO_2 배출량 데이터	

☞ '21년까지 농작물 품질생육이미지, 농업지식베이스, 가축행동, 축산품질, 어류개체행동 등 총 15종 구축·개방

참고 국내 분야별 데이터 구축 현황 분석

○ **(한국어)** 대화형 서비스의 기반 데이터 위주로 구축, 시멘틱 검색 및 기업 기술 경쟁력 확보의 원천이 되는 지식재산권 데이터 추가 확보 필요

○ **(영상·이미지)** 동작 추정 및 인체와 관련된 데이터를 다수 구축·보유, 최근 필요성이 증가하고 있는 영상 요약 서비스 관련 데이터 보완 필요

○ **(헬스케어)** 실제 국내 환자 분포를 고려한 주요 질병 위주 데이터가 구축되어 있으나, 심혈관질환 데이터 등 보완 필요

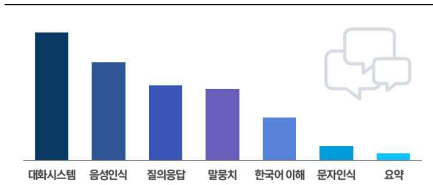
○ **(교통·물류)** 민간에서 수집이 어려운 장애물, 특수도로 차선, 포트홀 등 다양한 객체가 포함된 주행 관련 데이터 위주로 구축, 무인이동체·항만·물류 등 추가 필요

○ **(재난·안전·환경)** 이상행동CCTV, 공사현장, 화재 등 안전 관련 구축량이 상당 부분을 차지, 자연현상으로 인한 재난 관련 데이터 확보 필요

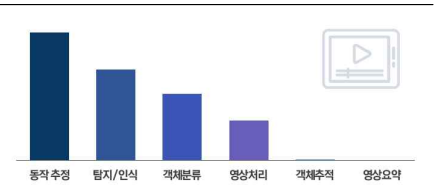
○ **(농·축·수산)** 최근 3년간 농작물 생산량 현황 비율*과 유사하게 농산물 데이터 구축, 수산의 경우 어류 위주의 데이터로 패류 등 추가 구축 필요

* **2017~2019** 노지 농작물 생산량 : 채소 > 미곡 > 과실 > 곡류기타 순

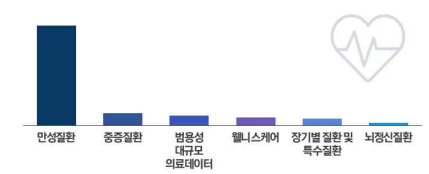
※ 출처 : 통계청, 통계로 본 농업의 구조 변화(2020), 발췌 정리



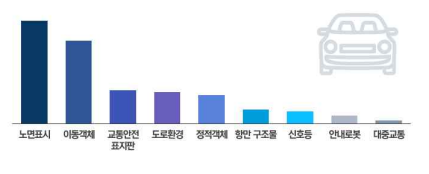
<한국어>



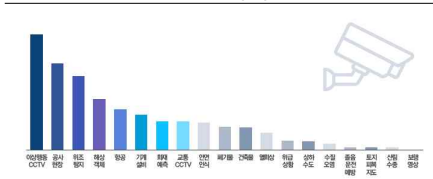
<영상·이미지>



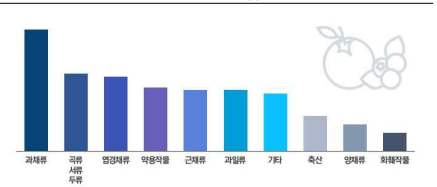
<헬스케어>



<교통·물류>



<재난·안전·환경>



<농축수산>

전략분야 2 : 인공지능 기반 혁신 촉진

1 (제 조) 다양한 **인공지능 솔루션 기업, 제조기업** 등이 **공정 최적화, 지능형 제조솔루션 개발** 등에 공통 활용 가능한 기반 데이터 구축

단기(~'23)	중기(~'25)
결함인식, 예지보전 등의 **기계·공정 관리** 및 **품질측정** 등을 지원하는 데이터	**수요 계획** 및 **예측, 제품설계**와 디자인, **자산 결함인식** 등을 지원하는 데이터로 확대

☞ 중소벤처기업부 'AI제조데이터 플랫폼(KAMP)'과 연계를 통한 전략적 협업 확대

< 제조 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	인공지능 기반 공정 최적화			지능형 제조·설비 관리 고도화	
제조관리	제조 품질검사 데이터, 공정 프로세스 제어 데이터			설비-데이터 연계형 공정 최적화 지원 데이터	
설비관리	이상 검출 및 고장 예측 데이터			설비 진단 및 수명예측 데이터 인간 변수 및 환경변화 기반 수명예측 데이터	

2 (로보틱스) **국내 기업·연구기관** 등의 **로봇 분야 기술력 증진, 용도·산업별 로봇 개발**과 **기능 고도화**를 지원할 수 있는 데이터 확보

단기(~'23)	중기(~'25)
로봇의 **인식·인지 능력 향상**을 지원하는 데이터 중심으로 확보	로봇의 **이동·모션** 등 **동작 제어**와 **인식·인지 능력 극대화**를 위한 데이터 구축

< 로보틱스 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	로봇 인식·인지 능력 향상			로봇 동작 제어	
제조·개인 서비스용	제품 정상·불량 판독 데이터 감정 상태를 반영한 엔터테인먼트 데이터			전천후 이동·배송 등을 지원하는 데이터, 자동화 공정 내 로봇 모션 조정 데이터	
전문 서비스용	상황·기능별 인지·해석용 데이터 징후 해석 및 진단을 위한 데이터			정밀진단 및 인식, 상황별 인지·판단능력 고도화를 위한 데이터	

3 (법 률) 국민의 **법률 분야 지식·정보 접근성 강화**와 분야별 **법률 서비스 제공 방식**의 **지능화** 등을 지원할 **기반 데이터** 축적

단기(~'23)	중기(~'25)
법률상담 음성·텍스트, 법률서식·용어·서적 등 **실생활 접근성**이 높은 데이터 구축	분야별 **법률상담 챗봇, 문서·계약서** 등 **자동 생성·검토, 선고예측** 지원 데이터 등 확보

< 법률 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	법률 기초·기반 정보 디지털화			인간-인공지능 협업 법률서비스 개발	
법률 정보	민·형사 및 행정 등 법령·판례 데이터 법률상담, 기초지식·용어 음성·텍스트 데이터			대화형 상담지원을 위한 분야별 법률지식 데이터	
법률서비스	분야별 문서·계약서 서식, 사례 등 데이터			문서·계약서 자동 생성 및 검토 지원 데이터, 사건별, 판례별 분석 데이터	

☞ 단계별로 확대되는 판결서 공개데이터(법원행정처) 등과 연계하여 민간 수요 기반 구축 추진

4 (교 육) 교육내용 전수(교사), 습득(학생) 및 교과 관리 등 **교수·학습 全 단계 지능화**와 **맞춤형 서비스 개발** 등을 위한 범용 데이터 구축

단기(~'23)	중기(~'25)
각종 **학습정보, 학습자 로그기록** 등 **인공지능 기반 교육환경 구축** 기반 데이터 확보	**다양한 교육과정 확보** 및 **학습자 맞춤형**의 **효율적 교육 추진**을 지원하는 데이터로 확장

< 교육 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	인공지능 기반 교육환경 지원			인공지능 기반 맞춤형 학습자·교육자 지원	
교 수 (Teaching)	지능형 튜터링 데이터, 비대면 교육용 데이터			실시간 학습활동 분석 데이터	
학 습 (Learning)	지능형 추천 등 학습 지원 데이터			지능형 학습자 관리 및 특수교육과정 지원 데이터	

☞ 교육부 'K-에듀 통합 플랫폼 사업' 데이터 등과 전략적으로 연계 구축

5 (금 융) 다양한 출처로부터 **투자정보**를 **생산·제공**하여 정보비대칭을 해소하거나, **서비스 개발·고도화**를 **지원**할 수 있는 데이터 확보

단기(~'23)	중기(~'25)
각종 **투자상담 사례, 전문용어, 시장정보** 등 **업무 자동화**를 지원하기 위한 데이터 확보	**금융 사기패턴 판별, 사이버 공격 탐지** 등 **리스크 관리, 준법 감시·대응** 지원 데이터로 확대

< 금융 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	지능형 금융 의사결정 지원서비스 개발			맞춤형 금융 분석지원 서비스 개발·고도화	
투자정보 분석	투자정보 추출·요약 데이터			국내외 투자정보 결합·분석 데이터	
금융소비자 보호	금융 전문용어 데이터			금융사기 모니터링(텍스트, 음성) 데이터	

6 (지식재산) 인공지능 기반 **지식재산 데이터**의 **전략적 활용**을 위한 **국내·외 인용정보** 및 **가치평가 연계정보** 등 **데이터 축적**

단기(~'23)	중기(~'25)
분야별 각종 특허 문헌, 특허-산업분류 정보 등 지식재산 관련 기초 데이터 구축	산업별 특허 데이터, R&D 연계 데이터 등 전략적 활용을 위한 데이터 구축

< 지식재산 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	각종 특허정보 데이터 확보			인공지능 기반 특허 관련 상담 및 지식재산 데이터 활용	
특허 상담	질문 인식, 매칭 답변 검색, 유사 질문 판별, 기계학습 기반 질의 의도 분석 등을 위한 기초 데이터			특허상담 유형 및 사례별 데이터	
유망기술 예측 등 지식재산전략수립	특허정보 - 산업분류정보 및 기술용어 연계 데이터 국내외 인용 정보 및 가치평가 정보 연계 데이터			각종 산업별 특허 데이터 특허정보-R&D 연계 데이터	

☞ 특허청 및 소속·유관기관 등 전문성을 갖춘 부처·기관이 과제 기획, 운영 등 총괄 추진

7 (문화·관광) 다양한 **지능형 문화·관광 서비스 구현**을 지원하는 **이미지, 지식DB**와 **증강현실·메타버스** 등 **첨단 서비스 적용**을 위한 데이터 확보

단기(~'23)	중기(~'25)
주요 관광지, 식당, 랜드마크 정보 등의 이미지·텍스트 데이터 등 확보	3D, 실감미디어, 메타버스 서비스 구현을 위한 기반데이터로 확대

< 문화·관광 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	지능형 서비스 도입 기반 마련			증강현실·메타버스 서비스 개발·확산	
문 화	문화유산, 전통문양 등 3D 이미지 데이터			한국적 콘텐츠 제작용 3D 이미지 데이터	
관 광	관광지식 DB, 랜드마크(문화유산), K-이미지(2D) 등			랜드마크 이미지(3D), K-이미지(3D), 3D 융합 데이터	

☞ 문화체육관광부 및 소속·유관기관 등 전문성을 갖춘 부처·기관이 과제 기획, 운영 등 총괄 추진

8 (스포츠) **선수 경기력 향상, 팀 전술 능력 향상**을 위한 **과학적 분석, 스포츠 저변 확대**를 위한 **서비스** 구현에 필요한 데이터 축적

단기(~'23)	중기(~'25)
피트니스, 스포츠 동작 관련 영상·이미지, 각종 스포츠 종목별 기초 데이터 구축	웨어러블 기기·센서 관련 데이터, 3D·모션 데이터 등 스포츠과학 지능화 지원 데이터 구축

< 스포츠 분야 데이터 로드맵(안) >

	2021	2022	2023	2024	2025
	신체동작 분석, 스포츠 경기 분석			판정·경기 분석·예측, 관련 분야 확산	
체 력	신체 동작 및 변화 분석 관련 영상·이미지			신체 동작 및 변화 분석 관련 영상·이미지	
운 동	동작·행동분류, 종목별 규칙·판정 기초 데이터			스포츠 종목별 전술패턴, 판정분석 데이터	

☞ 문화체육관광부 및 소속·유관기관 등 전문성을 갖춘 부처·기관이 과제 기획, 운영 등 총괄 추진

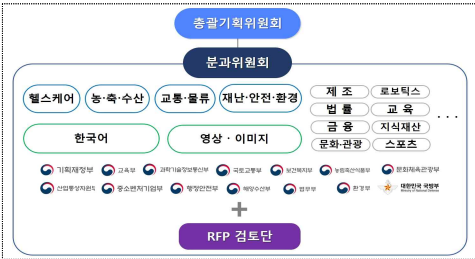
➡ 상기 예시 분야 외, **전략분야**를 **지속적**으로 **발굴·확장** 추진

3 로드맵 이행 체계를 갖추어 나갑니다

□ **주요 영역**별로 **전문성**을 **보유**한 **부처** 및 **전문기관**이 **과제기획, 추진방향 설정** 등 데이터 구축 전반에 **전략적**으로 **참여·주도**

○ 향후 **분야별 소관부처**가 해당 분야 **분과위** 및 **총괄기획위 참여** 추진

< 인공지능 학습용 데이터 과제기획위원회 구성(안) >



- **(총괄기획위)** 핵심 도메인 선정 및 구조화, 후보 데이터 선정, 지정 공모 대상 데이터 최종 검토 및 확정
- **(분과위원회)** 분야별 부처·전문기관, 산·학·연 전문가로 구성하고, 분야별 수요조사 검토, 후보 데이터 발굴, 우선순위 선정 등 수행

❶ **(사례1) 농·축산** ('22~) 해당 분야 전문성을 갖춘 **농식품부**가 **기획-구축-활용**을 **주관**하는 방식을 **도입·적용** 후 수요에 따라 **분야 확산**

※ '22년 협업예산 17대 과제 중 '디지털 기반 스마트 농업(농식품부, 과기정통부)' 과제를 통해 추진

< 농·축산 분야 '과기정통부-농식품부' 협력 방안('22~) >

- **(기획) 스마트농업 확산 거점**(혁신밸리, 노지 스마트농업 시범단지 등)과 연계하고, 농업인·기업, R&D 등 **현장 수요**를 반영한 **특화 데이터 구축 과제** 기획 협업
- **(구축) 농·축산 분야 전문기관**(농정원, 농진청 등)이 旣 보유한 **원천데이터**를 **제공**하고, 과제 수행기관 **선정**, **세부조정*** 및 **품질기준** 마련·적용 등 **참여·자문**

  * 데이터 구축 필요량, 수집 환경, 수집 기간 등을 농·축산 분야 전문기관과 협의·확정

→ **(활용)** 구축 및 품질검증이 완료된 데이터를 기반으로 **인공지능 경진대회**, **중소·스타트업 지원사업 연계, 스마트농업 확산** 등 **활용 활성화** 방안 공동 기획·추진

❷ **(사례2) 한국어** ('22~) **과기정통부-문체부**(국립국어원) 간 상호 **전략적 데이터 기획**과 **구축***을 위한 **협력 체계 운영** 및 **데이터 연계** 추진

* 음성자연어 분과위 등에서 문체부(국어원)를 포함한 데이터 기획 및 연계 협력 활용 방안 논의

□ 추후 「**데이터기본법**」에 따라 신설되는 '**국가데이터정책위원회**'의 분과로 '**인공지능 학습용 데이터 기획·활용**' 관련 **분과 신설** 검토

※ 분과 신설 시 과제기획위원회를 분과 내로 이관하여 범부처 협력의 제도적 근거 마련

2 구축 데이터의 활용성과 가치 증진 지원

1 쉽게 접근하고, 편리하게 활용할 수 있도록 합니다

□ (클라우드 토탈서비스) 데이터를 내려받지 않고도 **클라우드 환경**에서 **가공, 컴퓨팅자원 이용** 등이 가능한 **종합 지원체계** 구축·제공

○ 활용 빈도가 높은 **15종**의 데이터에 대해 **이용자 수요**를 토대로 **클라우드 기반 토탈서비스 시범 지원** 및 **점진적 확대 검토**('22~)

※ 장기적으로 AI Hub(스토리지)를 광주 데이터센터로 이관, 컴퓨팅자원 등 원스톱 제공('23~)

□ (데이터 접근성 개선) **데이터 분할·선택 제공, 딜리버리** 등 이용자 편의 증진을 위한 **데이터 접근성 개선방안** 시범 운영

데이터 분할제공	대용량의 데이터를 **필요한 만큼**만 내려 받아 **활용**할 수 있도록 **소규모·중규모**(1~20%) **버전 데이터** 제공 병행('22~) ※ 데이터를 내려 받기 전, 클라우드로 데이터 구성 등 확인 후 다운로드 규모 선택
객체선택 제공	**이용자가 원하는 객체만 선택**하여 다운로드 할 수 있도록 데이터에 포함된 **개별 객체를 검색**하여 **부분 다운로드** 제공('22~) ※ 객체 기반 데이터 검색을 통해 상세 데이터, 키워드, 연관 검색, 필터 기능 제공
데이터 딜리버리	네트워크 환경이 열악한 이용자도 **대규모 데이터**를 자유롭게 활용할 수 있도록 **저장매체 기반 '주문-대여 서비스***' 제공('22~, 자율주행 분야 시범) * 이용자가 활용하고자 하는 데이터(복수 가능)를 출력 포맷과 배송지를 선택하여 신청하면, 데이터셋을 생성한 후 저장장치에 복사하여 배송하는 방식

□ (데이터 활용·환류) **기 구축**된 **데이터**를 **서비스 개발**에 **활용**하는 과정에서 수집된 데이터를 **AI Hub**에 **축적·개방**하는 **선순환 체계** 마련

○ **'데이터 활용·환류'** 과제*를 신설하여 **'22년 시범도입** 후 확대 검토

* 기 구축된 AI Hub 데이터를 활용하여 인공지능 기반 제품·서비스를 개발·적용하고, 해당 서비스로 새로 생성·구축된 데이터를 다시 AI Hub에 공개

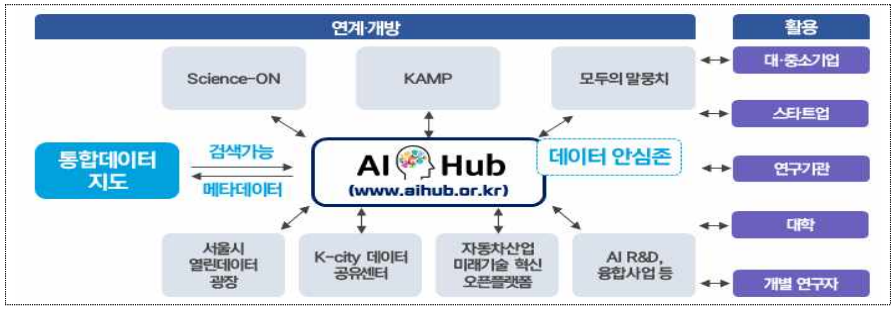
□ (AI Hub 연계·확장) AI Hub 데이터가 자유롭게 **수정·재가공**되어 **다양한 채널**로 **재배포**되도록 하여, **데이터**의 **지속적 발전** 및 **확산** 유도

※ AI Hub 데이터 구축기업의 동의를 기반으로 민간 자체 채널을 통해 개방하거나, AI Hub 업로드 및 민간 채널에 링크를 게시

○ 관련 **공공·지자체 포털** 및 **민간 제공 데이터***와의 **연계**도 **점진 확대**하여 인공지능 학습용 데이터 **검색·활용 편의 증진**

* **(LG CNS)** KorQuAD, **(카카오브레인)** KorNLI, KorSTS, **(네이버)** 클로바 콜 등, **(업스테이지·KAIST 등 10개 기업·기관)** KLUE 등

< 인공지능 학습용 데이터 연계·확장 개념도(안) >



□ (AI Hub 기능 고도화) **AI Hub**를 **이용자 간 상호 공유·경쟁**과 데이터 기반 **교육·실습**이 이루어지는 **종합 플랫폼**으로 고도화

○ 데이터 이용자의 **모델 성능 순위**를 **기록**하여 **경쟁**하는 **리더보드**를 **핵심데이터***를 중심으로 **상시 운영**

* 관심도 및 지속적 성능 개선이 요구·전망되는 태스크와 그에 맞는 데이터 선정 필요

※ 이미지넷 챌린지('10~'17)를 통해 시각지능 성능이 지속 개선되어 객체인식 정확도 98.8% 달성

○ 이용자 간 **성과** 및 **모델정보**를 **공유**할 수 있는 **커뮤니티 기능**을 추가하고, AI Hub 데이터 활용 **교육** 및 **실습 자료**를 **탑재**

□ (개인정보보호 강화) **AI Hub 데이터**를 **안심하고 활용**할 수 있도록 데이터 구축·활용 단계 전반의 **개인정보보호 조치** 추진

※ '개인정보 보호활용 기술 R&D 로드맵에 따라, 데이터 유형별 개인정보 탐지 등 기술개발 및 활용 지원('22~)

① **(가명·익명 처리 강화)** 데이터 **구축 단계**와 **개방 전**의 2단계에 걸쳐 가명·익명화 조치 **적용·검증**을 실시하고, **복수**의 **SW**로 **주기적 검증** 추진('21.12~)

② **(개인정보 신고제 운영)** AI Hub 데이터 내 **개인·민감정보** 발견 시 이를 **신고**할 수 있도록 하고, **확인된 사항**은 **즉시 조치** 및 **결과 피드백** 제공('22~)

③ **(가이드라인 마련·적용)** 데이터 **구축 수행기관·관계자** 누구나 쉽게 **참조**하고 **활용**할 수 있는 '데이터 유형별 개인정보 **가명·익명 처리 가이드라인**' 마련·배포('22)

2 다양한 정부 지원사업과 연계하여 활용 가치를 높여갑니다

□ (D.N.A. 지원사업간 연계) **디지털 뉴딜** 등을 통해 추진 중인 다양한 **데이터-클라우드-AI 지원사업**을 연계하여 제공하는 방안 도입

○ **국내 유망 인공지능 기업** 등을 대상으로 **복수**의 **정부 지원사업 참여 수요조사** 후, **공모**를 통해 **선정·지원**하는 **패키지형*** **지원** 추진('22~)

* '(가칭) Global AI Leaders' 프로젝트

< 단계별 지원 체계도 > * (예시) A기업, B기업

○ **데이터 바우처 지원사업**과 연계하여 중소·스타트업의 **AI Hub 데이터 재가공** 비용을 지원*하고, **재가공**된 데이터는 **AI Hub**에 **추가 공개**

* 기업별로 AI Hub 데이터를 개발하고자 하는 AI 솔루션 특성, 기존 데이터와의 정합성 등에 맞추어 새롭게 가공하여 활용하고자 하는 수요도 큰 상황

□ (지능정보화사업 연계) 데이터 구축·활용 등이 필요한 **지능정보화 사업***에 대해 **AI Hub 데이터 중복성 점검** 및 **우선 활용 권고**

* 「지능정보화기본법」에 따라 공공서비스의 지능정보화를 위해 국가가 추진하는 사업

※ AI Hub 데이터 활용도를 제고함과 동시에, 지능정보화사업 등에서는 과제 수행 상 필요한 인공지능 학습용 데이터 구축 시간·비용 절감 등 효율화 가능

○ 지능정보화사업의 **기획 단계**에서 도출된 **인공지능 학습용 데이터 수요**를 **AI Hub 데이터** 구축사업 **신규 과제기획**에 반영 추진

□ (타 사업 ← 데이터 제공) **인공지능 R&D, AI+X** 등 **인공지능 융합 사업** 추진 시 **AI Hub 데이터 활용 확대 지원** 및 **일부 의무화 검토**

○ **AI Hub**에서 제공하는 **데이터**와 **AI 모델**을 활용한 **R&D, 정책사업 과제**를 발굴·확산*하고, 필요에 따라 **활용**을 **의무화**하는 **과제 도입**

* (현재) 인공지능 그랜드 챌린지, 온라인 경진대회 → (향후) AI+X, AI R&D 등도 적용

3 믿고 쓸 수 있는 양질의 데이터로 만들어 갑니다

□ (선제적 품질관리) **데이터 구축 초기단계**부터 **인공지능 모델학습**을 **병행**하고, 오류 점검·보완 등을 위한 **사전검증** 및 **자문** 실시('22~)

데이터 구축 – 모델학습 병행	사전검증 및 자문
▪ 일정량 이상의 데이터 구축 후 **모델 학습 병행**을 통한 **활용성·품질 확인**	▪ 데이터 구축 중 오류발생 등이 우려되는 경우 **초기 데이터 사전검증** 추진
▪ 데이터 규격·분포 등 **설계오류, 라벨링 오류** 등을 **피드백·개선**	▪ 오류 수준이 기준초과 시 **즉시 보완 후 품질검증** 수행

□ (기 구축 데이터 고도화) **각종 변수**와 **환경**을 고려하여 **추가·보완**이 필요한 데이터의 경우 **신규 과제**를 통해 **고도화**하여 활용성 증진

○ **데이터 이용자**의 **피드백**과 **현황분석**에 기반한 **'기 구축 데이터 고도화' 과제 유형 신설** 추진('22)

※ 추가·보완 필요 데이터 현황조사 및 시범 과제 선정·추진 후 대상 확대

□ (맞춤형 데이터 구축) **데이터**의 **성격**과 **구축 난이도** 등을 고려한 **지원규모, 구축기간 다양화**를 통해 데이터 **품질·가치 극대화**

○ 과제 특성에 따른 **구축비용 지원 규모 차등화 방안*** 도입 추진

* **(현행)** 1종당 약 20억원 이내 → **(개선)** 대·중·소형 과제로 세분화 및 차등 지원

○ 연도별 과제 기획 시 **다년도**에 걸쳐 **구축이 필요한 과제**를 선정하여, 해당 과제군에 대한 **별도**의 **운영·평가체계**를 마련·적용('22~)

□ (데이터 신뢰성 확보) **데이터 수집-가공-개방** 각 과정의 **적법성, 투명성 점검** 및 **데이터 이력*** **관리 방안** 등 도입 추진('22~)

* 데이터 출처 및 수집·가공 방법, 적용가능 분야, 라이센스 정책 등

○ **AI Hub 데이터**의 **신뢰도**가 **이를 활용**한 **인공지능 솔루션**에 대한 **신뢰**로 **연결**될 수 있도록 개발자에 **데이터 출처 명시**를 권고

※ AI 솔루션 수요기업·기관에서 해당 AI의 신뢰성 및 윤리성 등을 사전 점검하기 위해 개발에 활용된 데이터의 이력 등을 함께 요구하는 사례 존재

4 데이터 구축에 함께한 분들의 성장을 지원합니다

□ (단계별 교육 활성화) **라벨러**부터 **관리자**까지 **맞춤형 교육 과정**을 **운영·확대**하여 데이터 구축 **전문성**을 높이고 **경력개발 기회** 제공

※ 필수 공통교육(인공지능 윤리, 개인정보 보호, 저작권 등)과 함께 수준별 교육 추진

< 인공지능 학습용 데이터 구축 참여인력 단계별 교육과정(안) >

구 분	교육 내용
일반라벨러	• **(입문)** 데이터 유형별 라벨링 기초, 라벨링 기본 가이드라인 등
	• **(기본)** 라벨링 도구 활용법, 오류분석, 작업속도 향상 기법 등
	• **(심화)** 유형별 고급 가공기법, 고난이도 분류 및 판별 기법 등
전문라벨러	• **(전문분야)** 특정 산업군(자율주행 등) 특화 라벨링
	• **(품질관리)** 유형별 품질관리 검사 절차 및 방법, 분석 방법 등
관리자	• **(프로젝트 관리)** 데이터 구축 생애주기별 프로젝트 관리, 운영 등

※ '21년말까지 온·오프라인의 라벨러 및 관리자 교육 3개 과정 등 총 1만명 수료 목표

○ 내실 있는 교육(매년 1만여명 규모)이 되도록 **이수평가**를 통한 **수료자 관리**와 **이수자 피드백 기반**의 **교육과정 지속 개선·확대** 추진

○ 참여인력 수준 향상을 감안하여 중·고급 과정인 **전문라벨러·관리자 과정 비중 점진 확대**

□ (他 교육과정 연계) 데이터 **구축·활용 저변**을 **확대**하고, 데이터 **구축 경험**을 관련 분야 **경력**으로 **발전**시키도록 **학교·직업교육** 등 연계('22~)

○ 분야별 **인공지능 학습용 데이터**를 활용한 **교육** 및 **실습과정**을 **개발**하여 **초·중·고등학교**를 중심으로 **보급·확산**

※ 과학창의재단(인공지능 교육교사 연구회 운영), SW산업협회(SW 전문강사 양성) 등과 협력하여, 의견수렴 및 교육·실습과정을 개발하고 확산 방안 마련

○ 데이터 구축 **참여이력**, 단계별 **교육 수료이력** 등을 **인공지능·데이터 인재양성** 및 **직업교육** 등과 **연계**하여 **경력개발 지원**

□ (크라우드소싱 적용 분야 확대) **단순 데이터 수집, 가공 작업**에서 **데이터 검수·품질관리** 등 **전문성**이 요구되는 분야로 **확대** 추진

※ 데이터 검수품질관리 관련 교육과정 도입·운영 및 일부 데이터 구축과제 적용 검토('22~)

Ⅴ. 추진일정

추진과제	관계부처	일정
1. 데이터 구축 · 개방 로드맵 수립 · 이행		
1 데이터 구축 · 개방 방향과 전략을 제시합니다		
• 인공지능 학습용 데이터 구축 로드맵 지속 보완	전 부처	'22~
• 세부 데이터 구축과제 기획	전 부처	'22~
2 시장에 필요한 데이터, 이렇게 구축 · 개방합니다		
• 한국어 분야 데이터 구축 확대	과기정통부, 문체부	'22~
• 영상·이미지 분야 데이터 구축 확대	과기정통부	'22~
• 헬스케어 분야 데이터 구축 확대	과기정통부, 복지부, 산업부	'22~
• 교통·물류 분야 데이터 구축 확대	과기정통부, 국토부	'22~
• 재난·안전·환경 분야 데이터 구축 확대	과기정통부, 행안부, 환경부	'22~
• 농·축·수산 분야 데이터 구축 확대	과기정통부, 농식품부, 해수부	'22~
• 주요산업 분야별(전략분야2) 데이터 구축 확대	전 부처	'22~
3 로드맵 이행 체계를 갖추어 나갑니다		
• 범부처 과제기획위원회 운영	전 부처	'22~
• 한국어 분야, 농·축산 분야 과제 공동기획 추진	과기정통부, 문체부, 농식품부	'22~
2. 구축 데이터의 활용성과 가치 증진 지원		
1 쉽게 접근하고, 편리하게 활용할 수 있도록 합니다.		
• 클라우드 토탈서비스 제공	과기정통부	'22.상~
• 데이터 접근성 개선방안 시범 운영	과기정통부	'22.상~
• 데이터 활용·환류 과제 신설·적용	과기정통부	'22~
• AI Hub의 민간채널과 연계·확장 강화	전 부처	'22.하~
• AI Hub 기능 고도화	과기정통부	'22.하~
• 개인정보보호 강화 대책 도입·추진	과기정통부, 개인정보위	'22.상~
2 다양한 정부 지원사업과 연계하여 활용 가치를 높여갑니다.		
• D.N.A. 지원사업간 연계방안 마련·추진	과기정통부	'22~
• 지능정보화사업과 데이터 구축활용 연계	전 부처	'22~
• 타 사업에 대한 AI Hub 데이터 제공 추진	전 부처	'22~
3 믿고 쓸 수 있는 양질의 데이터로 만들어 갑니다.		
• 선제적 품질관리체계 도입	과기정통부	'22.상~
• 기 구축 데이터 고도화 과제 신설·운영	과기정통부	'22~
• 데이터 특성별 맞춤형 데이터 구축방안 도입	과기정통부	'22~
• 데이터 신뢰성 확보 방안 도입	과기정통부	'22~
4 데이터 구축에 함께한 분들의 성장을 지원합니다.		
• 데이터 구축 참여자 수준·단계별 교육 활성화	과기정통부	'22~
• 他 교육과정과 연계 확대	과기정통부, 교육부, 고용부	'22~
• 크라우드소싱 적용 분야 확대	과기정통부	'22~

< 연도별 데이터 구축계획 및 소요예산(안) >

구분	'20년(추경)	'21년	'22년	'23년	'24년	'25년	합계
구축량(종)	150	190	310	220	220	210	**1,300**
소요예산(억원)	2,925	3,705	5,797	4,114	4,114	3,927	**24,582**

※ 관계부처 및 국회 등 협의에 따라 조정·변동 가능

■ 분야별 **관계부처** 및 **전문기관 협력** 하에 **인공지능 학습용 데이터 구축사업**(과기정통부)을 통해 '25년까지 **1,300종**의 **인공지능 학습용 데이터 구축·개방** 및 **활용 확산**을 추진