

데이터 댐을 위한 데이터 품질관리

김민준 한국지능정보사회진흥원 AI데이터추진단 책임연구원

임민성 한국지능정보사회진흥원 AI데이터추진단 선임연구원

1. 머리말

인공지능 학습용 데이터란 머신러닝, 딥러닝 등 AI 모델 학습에 활용되는 데이터를 총칭한다. 원본 데이터 및 활용 목적에 따라 원본 데이터에 표시 작업을 한 라벨링 데이터를 모두 인공지능 학습용 데이터라 하며, 원본데이터는 하나지만 라벨링 데이터는 사용 목적에 따라 다양한 형식으로 가공하여 사용할 수 있다.

인공지능 기술과 응용 서비스를 개발하려면 고품질 대용량 데이터가 반드시 필요하다. 그런데 인공지능 학습용 데이터를 구축하려면 비용, 시간 등의 현실적인 제약이 따른다. AI 학습용 데이터 구축 시 드는 비용은 데이터의 종류에 따라 다르지만, 평균 75%의 비용이 데이터 수집·정제·라벨링에 소요된다.

이러한 제약을 해결하기 위해 정부는 ‘인공지능 학습용 데이터 구축사업’을 추진하고 있다. 기업이 데이터 구축을 하는 데 소요되는 초기 비용과 시간을 줄이고 국가 AI 산업을 활성화하는 데 목표를 둔다. 2017년 시작한 ‘인공지능 학습용 데이터 구축사업’을 통해 현재까지 21종,

4650만 건의 데이터를 구축했고, 이를 AI Hub를 통해 개방하여 누구나 데이터를 다운로드하여 활용할 수 있다. 2020년에는 디지털 뉴딜 사업의 일환으로 ‘데이터 댐’ 구축사업을 본격화하여 2019년 대비 데이터 구축 종수는 2배 늘린 20종의 데이터(지정공모 10종, 자유공모 10종)를 구축하는 한편, 2020년 7월 추경을 통해 72개 과제 150종 세부 데이터를 구축 중이다.

이처럼 활용 가치가 높은 인공지능 학습용 데이터 구축·개방 노력이 이어지고 있지만, 여전히 인공지능 학습용 데이터의 양과 실효성에 관한 이슈가 있다. 구축 데이터의 품질은 곧 AI 서비스의 신뢰성과 안정성으로 이어지므로 데이터 품질은 매우 중요하다. 본고에서는 데이터 댐과 같은 대규모 데이터 구축사업 전반에서 핵심 요소라 할 수 있는 데이터 품질 관리 방법과 그 필요성에 관해 이야기하고자 한다.

2. 인공지능 학습용 데이터 구축사업 추진 경과

2.1 2017~2019년 인공지능 학습용 데이터 구축사업

정부는 2017년부터 산업적으로 활용 가치가

높은 법률, 특허, 일반상식, 한국인 안면 이미지 등 4개 분야에 대해 인공지능 추론·학습용 데이터인 지식베이스 구축사업을 추진했다. 이후 2018년에는 관광, 농업, 헬스케어, 음성, 대화, 멀티모달 등 세부 10개 분야에 대한 인공지능 학습용 데이터(지식베이스)를 구축하고 운영을 지원했다. 이어 2020년 1월에는 공공·민간에서 인공지능을 연구·개발하는 데 필수적인 인공지능 학습용 데이터, 인공지능 SW, 컴퓨팅 자원 등을 통합적으로 제공하는 포털인 AI Hub를 구축하여 데이터를 개방했다. AI Hub에서는 현재 인공지능 학습용 데이터 다운로드, 데이터 활용 시범 서비스, 공개 인공지능 SW 정보 제공 및 신청, 인공지능 컴퓨팅 자원 신청·승인 서비스 등을 제공하고 있다.

2019년부터는 인공지능 학습용 데이터 구축 사업을 본격적으로 추진했다. 해당 사업을 통해 10개 분야에 대해 인공지능 학습용 데이터 수집, 정제, 증강, 오류 검수, AI 시범 서비스를 제공했다. 구축된 인공지능 학습용 데이터는 AI Hub를 통해 2019년 12월 공개했다. 이와 함께 누구나 AI 응용제품·서비스 개발에 필요한 데이터를 자체적으로 구축하여 활용하고 공유할 수 있도록 각 분야 인공지능 학습용 데이터 구축 저작 도구(소스, 매뉴얼 등)를 함께 공개했다. AI Hub에서 텍스트, 이미지, 영상 등 원천데이터에 새로운 정보를 어노테이션, 태깅 등을 통해 가공할 수 있는 도구와 사전에 구축된 인공지능 학습용 데이터를 다운로드할 수 있다. 2017년부터 2019년까지 구축된 인공지능 학습용 데이터(지식베이스 포함) 구축사업으로 개방된 데이터를 <표 1>에 정리했다.

2.2 2020년 인공지능 학습용 데이터 구축사업

주제지정 인공지능 학습용 데이터 10종, 수요 기반 자유 공모 데이터 10종, 총 20종의 데이터 구축을 위해 데이터 수집, 데이터 정제, 데이터 증강, 원본 데이터 가공, 데이터 품질 검증을 수행하는 중이다. 2020년 10월 1차 사업으로 구축된 신규 인공지능 학습용 데이터(18종)를 AI Hub를 통해 1차 공개했다.

이에 더해 2020년 코로나19로 촉발된 경기침체에 대응하는 대규모 일자리 창출과 중소·벤처, 스타트업, 학계 등 민간의 인공지능 기술개발, 인공지능 응용서비스·제품 개발 촉진, 국가 인공지능산업 육성을 위해 10개 영역, 72개 과제(주제 지정 45개, 지역 6개, 자유 21개) 총 150종 데이터를 구축하여 AI Hub에 개방할 예정이다. [그림 1]과 [그림 2]에 2020년 추진 중인 인공지능 학습용 데이터의 종류를 나타냈다.

3. 데이터 품질 문제

지능정보산업 인프라 조성 사업으로 구축된 인공지능 학습용 데이터는 현재 AI Hub를 통해 무상으로 공급된다. 개방된 다양한 도메인의 데이터는 국내 산·학·연 기관의 연구개발, 서비스 개발을 포함하여 공공 및 민간에서 운영하는 각종 경진대회 등에서도 다양하게 활용되고 있다. 데이터 구축량이 증가하고 그에 따른 활용이 확대됨에 따라, 개방된 인공지능 학습용 데이터의 품질에 대한 의문과 관련 문제가 다방면으로 제기되고 있다.

인공지능 학습용 데이터의 품질 문제는 다음 4가지 유형으로 구분된다. 첫 번째는 데이터 다양성의 문제다. 이는 원시 데이터를 수집·구성할 때 특정 유형, 또는 종류에 치중하여 데이터 세트를 구축한 경우 발생할 수 있다. 데이터 다양

<표 1> 2017년~2019년 인공지능 학습용 데이터 개방 목록

분 야	지식베이스 구축 내용	예상 서비스
특 허	• 전기·전자분야의 특허기술 지식베이스 고도화·범위 확대 및 특허 청구항 분석(특성·패턴) 정보 추출을 통한 특허 지식베이스 구축	• 지능형 유사특허 검색 • 특허 분쟁예측 지원 등
법 률	• 교통사고, 퇴직금 등 생활관련 법령, 판례, 사례, 법률용어, 생활용어 등의 관계정보를 구조화한 지식 데이터	• 지능형 법률 검색 • 판결예측 지원 등
일반 상식	• 한국어 위키백과에서 추출한 객체정보와 관련 정보의 관계를 구조화한 지식데이터	• 인공지능 QA 서비스
한국형 이미지	• 다양한 각도와 환경에서 촬영한 한국 차량과 한국 음식 이미지 데이터	• 교통 관제 시스템 • 관광 안내 서비스
관 광	• 학습용 데이터(한/중/영/일 4개 국어 말뭉치 등) 및 관광 POI 데이터를 추출·가공하여 인공지능 추론·학습용 지식베이스 구축	• 지능형 관광챗봇 • 인공지능 관광가이드 서비스
농 업	• 농업 분야별 영농정보 및 농업상담·민원정보 등 농경데이터를 추출, 가공하여 지식베이스 구축	• 지능형 영농지원 상담 서비스 • 농업지원 프로그램 추천 서비스
헬스케어	• 안저(眼底) 이미지 분야 헬스케어 지식베이스 구축	• 안저이미지 기반 자동 선별 검사 서비스
멀티모달	• 인간 감정 및 의도 인지용 인공지능 개발을 위한 한국형 영상·음성·텍스트 융합 멀티모달 데이터	• 행동예측 AI • AI마케팅
음 성	• 음성인식 성능강화를 위한 한국어 대화음성DB 구축 * 그룹(2~10명)을 이루어 서로 자유롭게 대화하는 음성을 녹음하고 내용을 전사	• AIS피커 • 지능형로봇 • 자동회의록 작성 서비스 등
대 화	• 한국어 챗봇 성능 강화를 위한 한국어 자연어 대화 데이터	• AI챗봇 • AI상담사 등
기계독해	• 자연어 기반 한국어 지식·정답 추론 AI를 위한 한국어 기계독해(MRC) 데이터 20만 건 구축	• 지식추론 AI 및 컨설팅 AI서비스 등
번역 말뭉치	• 한국어 인공지능 번역 기술개발 및 성능 강화를 위한 한국어-영어 병렬 말뭉치 데이터 150만 문장 구축	• 한국어-영어 AI번역 시스템
사물 이미지	• 한국형 객체, 장소, 상황 인지기능 개발 및 성능강화를 위한 사물/거리/건물/랜드마크 등 사물 이미지 데이터 360만 장 구축	• 국내 사물인식 및 안내 서비스
글자체 이미지	• 한글 광학글자인식(OCR) 성능개선을 위한 한글 글자체(손글씨 및 인쇄체) 이미지 580만 장 구축	• 자동 한글 인식 서비스
인도 보행 영상	• 시각장애인/전동휠체어 등의 보행지원기술 개발을 위한 국내 인도·횡단보도 보행 영상 및 인도위 객체(사람, 자전거, 가로수, 펜스 등) 라벨링 데이터 200시간 구축	• 장애인 길안내 서비스 • 자율전동 휠체어
멀티모달 영상	• 감성인식 AI개발을 위해 동영상에서 인물의 표정, 음성, 발화 내용, 상황 등의 정보가 포함된 멀티모달 영상 데이터 100시간	• 인간의 감정, 상황 이해 로봇
사람 동작 영상	• 사람의 동작·자세·행동 인식기술 개발을 위해 다양한 조건에서 사람 동작 영상 데이터 50만 건 구축	• 가정, 공공시설 등의 서비스 로봇
안면 이미지	• 다양한 각도, 조도 등의 환경하에서 안면인식·식별 성능 강화를 위한 한국인 얼굴 이미지 데이터 1,600만 장 구축	• 한국인 얼굴 인식 AI시스템, 범죄자 탐지
위험물 이미지	• 위험물·도구 자동판별 기술개발 및 성능개선을 위한 위험물, 범죄 도구, 반입금지물품 등의 X-ray 이미지 48만 장 구축	• 공항, 항만, 철도, 주요 시설 AI보안 검사 시스템
질병 진단 이미지	• 안전질환 이미지, 유방암 엑스레이 이미지 및 진단결과(질환 및 정상)를 라벨링한 질병진단 이미지 데이터 3만 장 구축	• 주요질환 자동검진 AI의사
이상 행동 영상	• CCTV 영상에서 사람 등의 이상행동 지능형 탐지기술개발 및 성능 강화를 위한 이상행동 영상 데이터 700시간 구축	• 지능형 CCTV • AI 안전감시 시스템

성의 문제는 대량의 데이터 품질 검수를 샘플링 방식으로 수행할 때 발생한다. 특히 데이터 설계 시 다양한 클래스를 포함할 때, 클래스 간 구성의 차이로 인해 이와 같은 문제가 발생할 가능성

이 크다.

두 번째 품질 문제는 인공지능 학습용 데이터의 레이블 정확도 문제다. 레이블 데이터는 인공지능 학습용 데이터 중 가장 중요한 요소이자,

 문서 요약 텍스트 AI 데이터 도서, SNS, 논문 등 분야별 원본 텍스트 데이터 35만 건 이상과 원문의 핵심 내용을 적절히 요약한 텍스트 데이터	 전문분야 한-영 말뭉치 AI 데이터 전문분야를 선정하고 분야별로 적정하게 배분한 한국어-영어 병렬 말뭉치 150만 건 이상
 대용량 동영상 콘텐츠 AI 데이터 다양한 분야별 원천 비디오 데이터에 객체, 객체위치, 상황 등이 결합된 학습용 영상 데이터 50시간 이상	 랜드마크 이미지 AI 데이터 전국 50개 도시의 랜드마크(국내 공공기관, 주요건물, 관광명소, 편의시설 등) 이미지 총 500만 장 이상
 딥페이크 변조영상 AI 데이터 성별, 나이등을 균등하게 고려하여 제작된 원본 영상을 기반으로 합성(변조)된 동영상 데이터 20만 개 이상	 자율주행드론 비행영상 AI 데이터 다양한 비행환경(시간, 장소 등)에서 획득된 영상과 장애물 등의 객체가 레이블링 된 영상 총 300시간 이상
 수어(手語)영상 AI 데이터 일상생활 및 특정상황에 사용되는 단어(4,000단어 이상) 및 문장(2,000문장 이상)을 표현하는 동영상 20만 개	 한국인 대화음성 AI 데이터 성별, 지역 등 분야별 원본 음성 데이터 4,000시간 이상과 음성을 문자로 변환한 텍스트 데이터 400만 문장 포함
 시각정보기반 질의 응답 AI 데이터 다양한 객체, 상황 등을 표현하는 이미지 115만 장 및 질문-답변 데이터 690만 개	 사람 신체·자세 3D AI 데이터 70여 개 이상 사람의 행동 및 동작으로 3D 모델 50만 건, 2D 이미지 200만 장 이상
 질병진단 이미지 AI 데이터 4,000명에서 추출된 유방암 병리 이미지 10만 건, 부비동(DICOM) 데이터 8,000건	 도로주행영상 AI 데이터 실제로 주행 데이터 Use-Case 기반 자율주행원천데이터 수집, 총 학습용 데이터 60만 5천 프레임 이상
 도로환경 파노라마 이미지 AI 데이터 도심지파노라마(3,400km, 339,910장, 94시간 규모), 평면영상 (2,711,280장, 753시간 규모)	 치매진단 뇌파영상 AI 데이터 치매관련 의료 영상 데이터와 진단정보를 결합한 AI 데이터 28만 건
 피트니스 자세 이미지 AI 데이터 200,000개의 비디오로 구성하며, 각 운동 종목당 최소 2,400개의 비디오 포함	 감성 대화 말뭉치 AI 데이터 우울증 등 심리 장애로 인한 사회문제 해결을 위한 감성대화 코퍼스 데이터 270,000문장, 발화 음성 10,000건
 K-Fashion 이미지 AI 데이터 저작권 문제가 해결된 패션 이미지의 패션요소 정보를 어노테이션한 이미지 120만 장 이상	 위성영상 객체판독 AI 데이터 광학 및 레이더 영상을 이용한 5종의 AI 데이터 120만 건 구축
 한국인 재식별 이미지 AI 데이터 대한민국의 실내/외 구축된 공공 CCTV 환경을 고려한 한국인(1,000명) 재식별 데이터 세트 400만 장	 사람 신체·자세 3D AI 데이터 치아 및 치주질환 진단과 치료계획 수립을 위한 파노라마 5,000장(5,000명), CBCT 200,000장(2,000명)

[그림 1] 2020년 1차 구축 인공지능 학습용 데이터(20종)

 자연어(97개) 음성인식 고도화 등을 위한 음성 및 말뭉치 데이터 등 32종	 미디어(47개) 영상내 음성·글자인식, 아외소음환경, 스포츠(축구, 농구, 골프) 사람동작 데이터 등 11종
 헬스케어(97개) 암, 치매, 피부질환 진단을 위한 의료 영상 데이터, 진료 및 건강 데이터 등 20종	 안전(67개) 안전감시 및 안전활동 자동화를 위한 영상 데이터, 노후시설물 이미지 데이터 등 16종
 자율주행(57개) 자율주행기술발전을 위한 차선·횡단보도, 신호등·도로표지판 데이터 등 14종	 기타(57개) 소상공인, 제조 분야의 업무 자동화 및 비용 절감을 위한 영상, 이미지 데이터 등 8종
 농축수산(47개) 농·수산물 지능형 로봇 개발과 작물 질병·해충 진단을 위한 영상 및 이미지 데이터 등 12종	 지역분야(67개) 지역(지자체) 수요를 반영한 자유주제 AI 학습용 데이터 선정 및 구축 추진
 국토환경(37개) 환경 오염 모니터링과 기후 변화 예측을 위한 이미지 데이터 및 위급상황데이터 등 10종	 자유(217개) 최신 시장 수요 등을 반영한 자유주제 AI 학습용 데이터 선정 및 추진

[그림 2] 2020년 2차(추경) 구축 인공지능 학습용 데이터(150종)

일반적으로 가장 오랜 공수가 필요하다. 실제 레이블 데이터를 만드는 가공 과정, 또는 가공 중 오류를 수정하는 검수 과정이 주로 인력에 의해 진행되므로, 정확하지 않은 레이블 정보가 입력되어 발생할 수 있다.

앞서 언급한 레이블 정확도와 마찬가지로 비디오 또는 음성 데이터에 주로 발생하는 문제로, 레이블 정보에 포함된 시점정보와 비디오 또는 음성의 재생정보가 일치하지 않는 시점 불일치의 문제가 있다. 레이블 데이터에는 특정 시점에서 해당 이벤트가 발생한다고 기술되어 있으나 실제로는 이벤트가 발생하지 않거나, 다른 시점에서 발생하는 경우 해당 문제가 있다고 할 수 있다. 이러한 시점 불일치의 문제 역시 인력을 통해 데이터를 가공하거나 검수하기 때문에 발생한다고 할 수 있다.

마지막으로, 데이터 구축 목적과 다른 용도로 인공지능 학습용 데이터를 이용하려는 경우 발생하는 활용 목적 불일치의 문제가 있다. 인공지능 모델을 학습하여 서비스 개발에 이용하는 목적으로 데이터를 구축했음에도, 이와 다른 목적으로 데이터를 활용하고자 할 때 발생하는 문제다.

4. 인공지능 학습용 데이터 품질관리 및 표준화 방안

인공지능 학습용 데이터를 활용하는 과정에서 발생할 수 있는 문제들은 다음 사유로 인해 발생할 수 있다.

- 품질기준 미흡
- 사후 유지보수 체계 부재
- 도메인 전문성 한계
- 구축 수행기관 책임감 담보 미흡

품질기준이 명확하지 않거나, 기준이 있음에도

이를 준수하지 않을 경우 품질 문제가 발생할 수 있다. 또한, 품질기준이 레이블 오류 점검 위주로 설정된 경우에는 데이터 다양성 측면이나, 유효성 측면의 검증을 수행할 수 없으므로 문제가 발생할 여지가 크다. 품질 문제가 발생하였을 경우 신속하게 문제에 대응할 수 있는 관리체계가 부재한 것도 문제를 확대할 수 있는 요인이다.

데이터가 특정 도메인(의료, 농업, 자율주행 등)과 관련성이 높을 경우 해당 분야에 대한 전문적인 검증에 약점이 있을 수 있다. 특히, 데이터 활용기업이 직접 데이터 검증에 참여할 수 없는 체계로 인해 활용 단계에서 다양한 품질 문제에 직면할 수 있다.

마지막으로, 실제 인공지능 학습용 데이터 구축을 담당하는 수행기관에 책임감을 담보할 방법이 부재한 것도 품질 문제 발생 가능성을 높이는 한 가지 요인이 된다. 데이터 구축 수행기관이 데이터 품질관리에 대한 인식과 역량을 갖추고, 품질 검증 결과에 따른 인센티브나 제재 방안이 반드시 필요하다. <표 2>는 다양한 데이터 유형에 공통적으로 적용할 수 있는 품질 검증 항목 및 기준을 제시한다.

공통 품질관리 기준에는 인공지능 학습용 데이터를 구축하는 절차에 대한 기준과 원시 데이터 및 레이블 데이터에 대한 품질 검증 기준을 포함했다. 기준 선정 시 사업 추진 초기 단계부터 데이터 구축 수행기관과 지속적으로 합의를 진행했다. 또한, 이렇게 선정된 기준 중 데이터 획득, 가공, 검수, 활용의 전 주기를 포괄한 구축 절차에 대한 품질 검증 대상, 검증 방안에 대한 단체 표준안을 제시하고자 한다.

부가적으로 AI Hub에 개방된 데이터의 활용 목적 불일치의 문제가 발생하지 않도록 다양한 방안을 마련하고 있다. 각 데이터에 대한 소개자

<표 4> 인공지능 학습용 데이터 공통 품질관리 항목

검증 대상		주요내용	검증 기준
품질관리 절차	절차 준수율	공통 기준상 품질 관리 절차 준수율	• 문제 정의 • 데이터 수집 • 정제·가공 • 데이터 검수 • 데이터 활용
원시 데이터	기계학습 적합성	획득 데이터가 기계학습으로 문제를 해결하는 데 적합하게 구성되어 있는지 점검	• 다양성 • 신뢰성 • 공정성 • 충분성 • 균일성 • 사실성
레이블 데이터	레이블 정확성	데이터 레이블이 빠짐없이 정확하고 일관성 있게 작업됐는지 검사	• 어노테이션의 기능목적 적합성 • 객체 분류의 명확성 • 어노테이션 특성 정보의 포괄성
	기계학습 유효성	AI 알고리즘을 훈련시켰을 때 목표로 했던 수준의 성능 달성 여부	• 학습 유효성

료를 보강하고, 데이터를 활용한 서비스 수준도 높여 이용자가 쉽게 데이터를 이해할 수 있도록 한다. 데이터 댐의 핵심 사업 중 하나인 인공지능 학습용 데이터 구축사업은 구축 준비단계부터 활용단계까지 전체 과정에 대해 빈틈 없는 품질 관리 기준을 마련하고자 노력하고 있다.

5. 맺음말

2020년 7월부터 국내 경기를 부양하고 사회 전반의 격차를 해소하기 위해 한국형 뉴딜 사업이 추진되고 있다. 그중에서도 가장 많은 예산이 투입되는 디지털 뉴딜을 이끄는 핵심 사업 중 하나인 데이터 댐 구축사업에 사회적인 관심이 집중되고 있다. 많은 예산이 투입되는 사업인 만큼 사업 수행 성공을 위해 양적인 만족도를 넘어 질적으로도 충실도를 확보해야 할 것이다.

본고에서는 인공지능 학습용 데이터 구축사업

을 통해 구축되는 데이터의 품질 문제에는 어떤 것들이 있으며, 이러한 품질 문제를 해결하기 위해 추진하는 노력은 무엇인지 기술했다. 이에 따라 단순히 데이터 구축 과정뿐 아니라 구축 절차에 제시된 설계, 구축, 검수, 활용의 전 단계에 걸친 데이터 공통 품질 관리 기준을 제시했다. 현재 공통 품질 관리 기준을 바탕으로 단체 표준화 기구를 통해 단체 표준 작업이 진행 중이며, 해당 성과는 이후 국내 및 국제 표준화 성과로 이어질 것으로 기대된다.

본고에서 제시한 공통 품질 관리 기준과 표준화 방안은 인공지능 학습용 데이터 구축사업의 성과물이 높은 품질을 달성할 수 있도록 지속해서 적용, 관리돼야 한다. 인공지능 기술 발전에 따라 공통 품질 관리 기준을 적합하게 적응시켜 나가는 한편, 인공지능 학습용 데이터의 국제표준화를 선도하고 우리나라가 인공지능 강국으로 도약하기 위해 더욱 노력할 계획이다. TTA

참고문헌

- [1] AI Hub, www.aihub.or.kr
- [2] 한국정보화진흥원(2020.3), '인공지능 학습용 데이터 구축사업' 공모안내서
- [3] 한국정보화진흥원(2020.7), '인공지능 학습용 데이터 구축사업(2차)' 공모안내서
- [4] 한국정보화진흥원(2020.11) IT&Future Strategy 제7호 'AI 학습용 데이터 사업의 실효성 향상을 위한 정책 방향'

주요 용어 풀이

- **관심지점(Point of Interest)**: 특정 지점, 시설물의 좌표 데이터
- **기계독해(Machine Reading Comprehension)**: AI가 제시된 문단(문장의 집합) 안에서 정답을 추론하는 딥러닝 기반 기술
- **레이블 데이터(Labeled data)**: 원시 데이터를 정제·가공을 통해 인공지능이 학습 진행할 수 있도록 설명정보를 추가한 데이터
- **어노테이션**: 인공지능 학습용 데이터 가공 시 설명정보를 표현하는 방식
- **원시데이터(Raw data)**: 데이터 설계 이후 수집되는 데이터 자체를 의미하며, 별도 정제나 가공 과정을 거치지 않은 데이터
- **인공지능 학습용 데이터**: 인공지능이 학습을 통해 성능을 향상시키는 데 활용되는 데이터
- **지식베이스(Knowledge-base)**: 기계가 이해하고 학습할 수 있도록 기존의 데이터에 태그 삽입, 용어사전, 의미관계 등으로 구조화하여 정제된 인공지능용 추론·학습용 데이터