

| 메타데이터 정보 (다중기입가능) | 분야 | 데이터 유형 ¹⁾ | 구축 데이터량 | 원천데이터 형식 ²⁾ | 라벨링 형식 ³⁾ | 라벨링 유형 ⁴⁾ |
|-------------------------|----------------------|--|----------|---------------------------|-----------------------------|----------------------|
| | 영상이미지 | 이미지 | 100,000 | png | json | 바운딩박스 (이미지) |
| | 데이터 출처 ⁵⁾ | 데이터 구축년도 | 구축기관(총괄) | 가공기관 | 검수기관 | |
| | 자체수집 및 제공 협약 | 2021년 | 동양시스템즈 | (주)쇼우테크 | TTA | |
| | 데이터 문의처 | 기관명 | 문의담당자명 | 전화번호 (유선전화번호기입) | 메일주소 | |
| | | (주)쇼우테크 | 황강연 | 055-323-3169 | gyhwang@ ishowtech.co.kr | |
| | 데이터 소개 | 공공기관 및 민간에서 사용하는 다양한 정보제공형 서식과 일상생활에서 많이 사용하는 단어에 대한 손글씨를 인식할 수 있는 모델 개발에 필요한 손글씨 OCR 데이터 구축 | | | | |
| | 주요키워드 | 인공지능, OCR, handwriting, AI, Paper, Tablet, 문자인식, 손글씨 | | | | |
| 카테고리 정의서 | | 첨부의 카테고리 정의서 엑셀파일에 데이터카테고리 작성하여 제출(예시참고) | | | | |

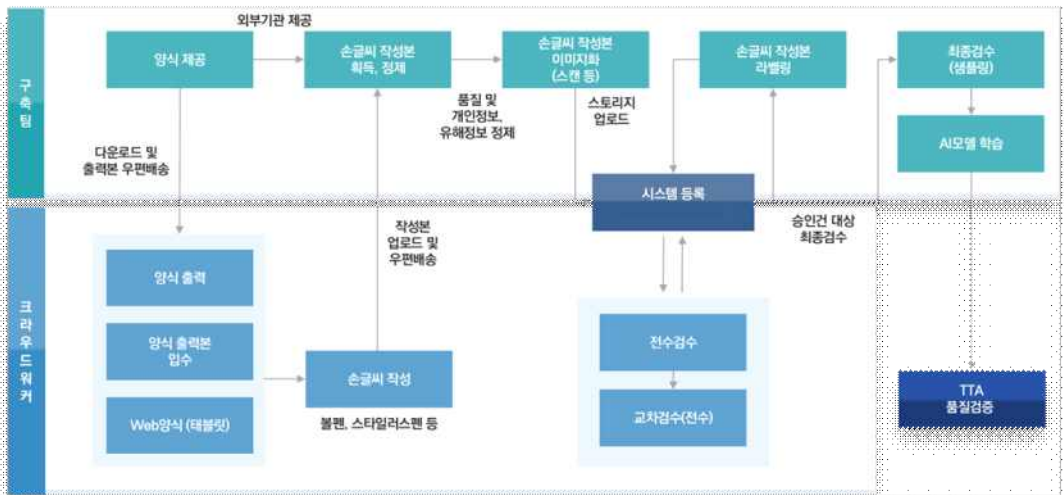
1) 텍스트, 오디오, 이미지, 비디오,

2) txt, jpg,.....

3) json, csv,.....

4) 내용요약(텍스트), 번역(자연어), 질의응답(자연어), 바운딩박스(이미지/동영상), 키폰트(이미지/동영상), 세그멘테이션(이미지/동영상), 전자(음성)

5) 4대 언론기사, 자체 수집,.....

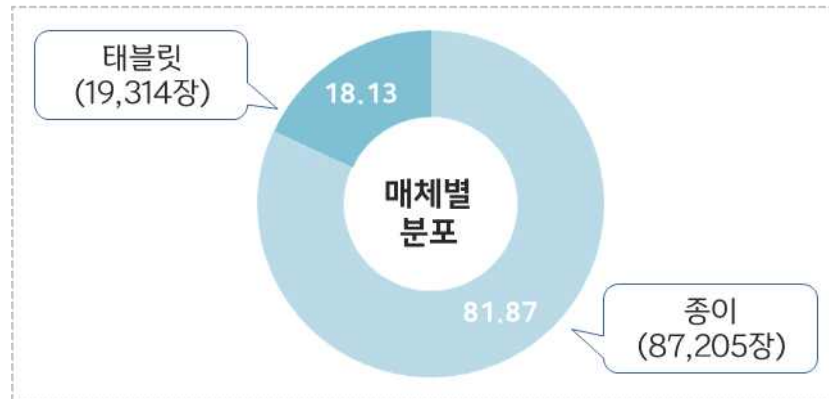
| 데이터셋명 | 국문영문 | 대용량 손글씨 AI 데이터셋 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-------------------------|--|---|---------|--------------|--------|---|------|------|----|----------|-----------------|--------|--------------|----|-------------------|----------|--------|--------------|----|------|--------|-------|----|----------|---------|---------|--------|----|------------------------------------|-----|----------|---------|--------|--------|--------|---|----------|------|---------|--------|--------|---|----|--|------|--|--|
| | 문 | Large capacity handwriting artificial AI Training Dataset | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 구축목적 | 다양한 내용, 성별, 연령대에서 작성하는 손글씨 데이터 원문으로부터 손글씨 OCR 인식이 가능하도록 인공지능을 훈련하기 위한 데이터셋 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 활용서비스 | 현재까지 공공기관, 학교 등에서 손글씨로 제공받을 수 밖에 없는 서식, 시험답안지 등에 기재된 손글씨 데이터를 인식하고 텍스트로 제공하는 손글씨 OCR 인식 서비스 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 소개 | 종이문서와 태블릿 등 다양한 형태로 작성된 손글씨 데이터 수집 및 OCR 인식을 위한 AI 데이터셋으로, 일상생활에서 많이 사용하는 단어를 중심으로 개인정보 등의 민감정보를 제거한 원천 데이터 확보 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 데이터셋 통계 (구축 규모 및 분포) | 1. 데이터 구축 규모 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | <table><tr><th colspan="2">항목</th><th>데이터량</th><th>획득처</th><th>획득유형</th><th>획득방법</th></tr><tr><td rowspan="4">종이</td><td rowspan="3">자유 필사</td><td>공무원 목민심서 필사본</td><td>1,000장</td><td>창원시청 김해시청</td><td>스캔</td><td rowspan="3">1) 원본 수령 2) 스캔</td></tr><tr><td>대학 시험답안지</td><td>3,000장</td><td>부산경상대 동의대</td><td>스캔</td></tr><tr><td>노트필기</td><td>3,000장</td><td>부산경상대</td><td>스캔</td></tr><tr><td>정보 제공</td><td>자체 워크시트</td><td>30,000장</td><td>클라우드소싱</td><td>스캔</td><td>1) 가이드 배포 2) 작성본 우편 수령 3) 스캔</td></tr><tr><td rowspan="2">태블릿</td><td>자유 필사</td><td>자체 워크시트</td><td>4,500장</td><td>클라우드소싱</td><td>e-Form</td><td>1) 가이드 배포 2) e-Form 작성 3) 플랫폼 업로드</td></tr><tr><td>정보 제공</td><td>민원양식</td><td>13,500장</td><td>클라우드소싱</td><td>e-Form</td><td>1) 가이드 배포 2) e-Form 작성 3) 플랫폼 업로드</td></tr><tr><td colspan="2">합계</td><td>10만장</td><td colspan="3"></td></tr></table> | | 항목 | | 데이터량 | 획득처 | 획득유형 | 획득방법 | 종이 | 자유 필사 | 공무원 목민심서 필사본 | 1,000장 | 창원시청 김해시청 | 스캔 | 1) 원본 수령 2) 스캔 | 대학 시험답안지 | 3,000장 | 부산경상대 동의대 | 스캔 | 노트필기 | 3,000장 | 부산경상대 | 스캔 | 정보 제공 | 자체 워크시트 | 30,000장 | 클라우드소싱 | 스캔 | 1) 가이드 배포 2) 작성본 우편 수령 3) 스캔 | 태블릿 | 자유 필사 | 자체 워크시트 | 4,500장 | 클라우드소싱 | e-Form | 1) 가이드 배포 2) e-Form 작성 3) 플랫폼 업로드 | 정보 제공 | 민원양식 | 13,500장 | 클라우드소싱 | e-Form | 1) 가이드 배포 2) e-Form 작성 3) 플랫폼 업로드 | 합계 | | 10만장 | | |
| 항목 | | 데이터량 | 획득처 | 획득유형 | 획득방법 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 종이 | 자유 필사 | 공무원 목민심서 필사본 | 1,000장 | 창원시청 김해시청 | 스캔 | 1) 원본 수령 2) 스캔 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | 대학 시험답안지 | 3,000장 | 부산경상대 동의대 | 스캔 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | 노트필기 | 3,000장 | 부산경상대 | 스캔 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | 정보 제공 | 자체 워크시트 | 30,000장 | 클라우드소싱 | 스캔 | 1) 가이드 배포 2) 작성본 우편 수령 3) 스캔 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 태블릿 | 자유 필사 | 자체 워크시트 | 4,500장 | 클라우드소싱 | e-Form | 1) 가이드 배포 2) e-Form 작성 3) 플랫폼 업로드 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | 정보 제공 | 민원양식 | 13,500장 | 클라우드소싱 | e-Form | 1) 가이드 배포 2) e-Form 작성 3) 플랫폼 업로드 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 합계 | | 10만장 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | ※ 이미지 기준 10만장, 이미지 당 40단어 이상 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

2. 데이터 분포

가. 매체별 분포

- 종이 : 종이문서에 기재한 손글씨(스캐닝하여 디지털화)
- 태블릿 : 태블릿 기기에 펜을 이용하여 기재한 손글씨

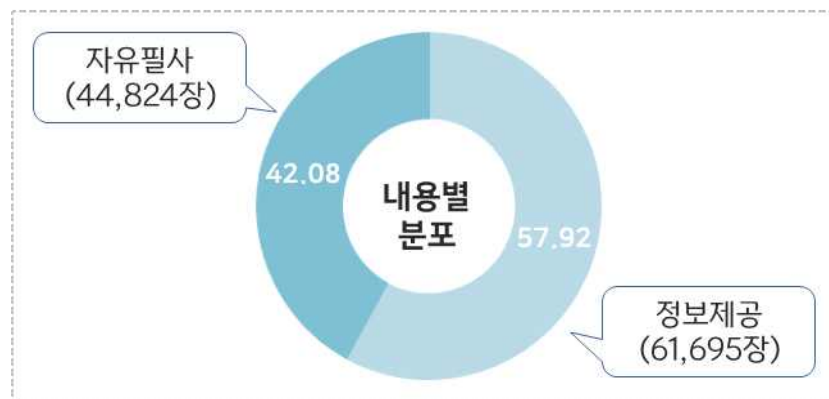
| 항목 | 데이터량 | 비율 |
|-----|----------|--------|
| 종이 | 87,205장 | 81.87% |
| 태블릿 | 19,314장 | 18.13% |
| 합계 | 106,519장 | |



나. 내용별 분포

- 정보제공형 : 공공기관 및 산업에서 활용하는 서식
- 자유필사형 : 자유형태의 손글씨 데이터 및 목민심서 필사본, 시험답안지, 노트필기류 등

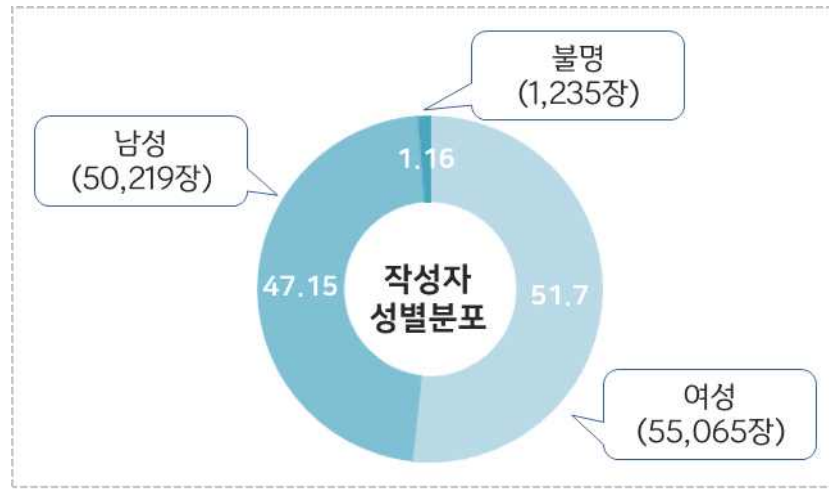
| 항목 | 데이터량 | 비율 |
|------|----------|--------|
| 정보제공 | 61,695장 | 57.92% |
| 자유필사 | 44,824장 | 42.08% |
| 합계 | 106,519장 | |



다. 작성자 성별 분포

- 작성자의 성별 분포 50:50으로 수집
- 성별을 판별할 수 없는 데이터(공무원 목민심서 필사본은 “불명”으로 수집)

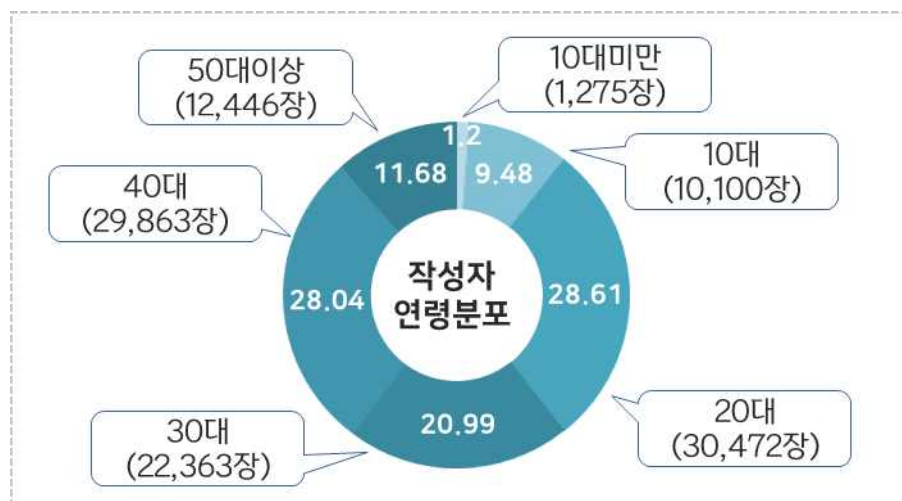
| 항목 | 데이터량 | 비율 |
|----|----------|--------|
| 여성 | 55,065장 | 51.70% |
| 남성 | 50,219장 | 47.15% |
| 불명 | 1,235장 | 1.16% |
| 합계 | 106,519장 | |



라. 작성자 연령 분포

– 10대 ~ 50대까지 다양한 연령대를 대상으로 수집

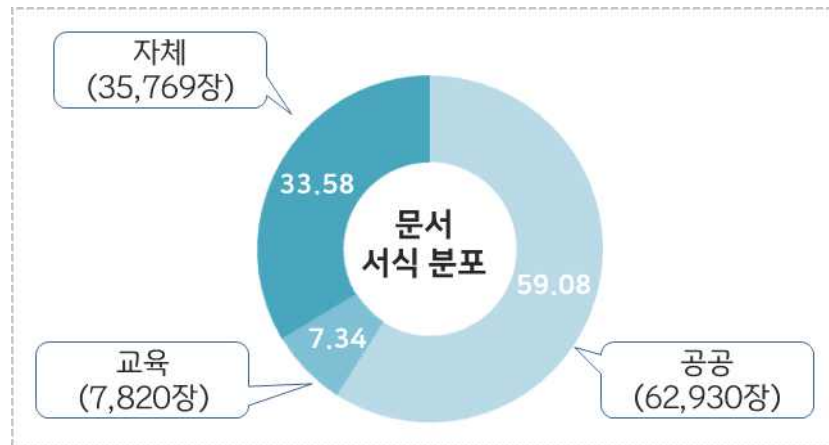
| 항목 | 데이터량 | 비율 |
|-------------|----------|--------|
| 10대 미만(~9) | 1,275장 | 1.20% |
| 10대(10~19) | 10,100장 | 9.48% |
| 20대(20~29) | 30,472장 | 28.61% |
| 30대(30~39) | 22,363장 | 20.99% |
| 40대(40~49) | 29,863장 | 28.04% |
| 50대 이상(50~) | 12,446장 | 11.68% |
| 합계 | 106,519장 | |



마. 문서 서식 분포

- 공공 : 공공기관에서 사용하는 민원서식 및 목민심서 필사본 등
- 교육 : 교육기관에서 수집한 논술형 시험답안지 및 노트필기류 등
- 자체 : 자체 제작 워크시트

| 항목 | 데이터량 | 비율 |
|----|----------|--------|
| 공공 | 62,930장 | 59.08% |
| 교육 | 7,820장 | 7.34% |
| 자체 | 35,769장 | 33.58% |
| 합계 | 106,519장 | |



1. 대표도면

데이터셋 구성

| 파일명 | |
|-------------------------------|------|
| IMG_OCR_53_4PR_17040 | |
| 데이터셋 타입 | 이미지 |
| 데이터셋 카테고리 | OCR |
| 기록매체 유형 | 종이 |
| 작성내용 | 자유필사 |
| 수집장소 | 자체 |
| 작성자 성별 | 여성 |
| 작성자 연령 | 36 |
| png json | |

| No. | 속성명 | 항목설명 | Type | 필수 여부 | 작성예시 |
|------|-------------------------------|-----------------------|--------|----------|-----------------------------------|
| 2-15 | Images.writer_sex | 작성자 성별 | number | 필수 | 0: 여성, 1: 남성, 2: 불명 |
| 2-16 | Images.data_captured | 이미지 생성 일자 | string | 필수 | yyyy.mm.dd HH:MM:SS |
| 2-17 | images.written_content | 작성내용 | number | 필수 | 0: 정보제공, 1: 자유필사, 2: 기타 |
| 3 | Annotation | 어노테이션방식 | | | |
| 3-1 | Annotation.object.recognition | 개체 인식 | number | 필수 | 0 : 바운딩박스 1 : 폴리곤 |
| 3-2 | Annotation.text_language | 라벨링 텍스트 언어 | number | 필수 | 0: 한글, 1: 한자, 2: 영어, 3: 숫자, 4: 기타 |
| 4 | BBox | 바운딩박스 어노테이션 구조 | | | |
| 4-1 | BBox[.id | 바운딩박스 식별자 | number | 필수 | BBX_001(분류_순번) |
| 4-2 | BBox[.text | 바운딩박스 내 텍스트 | string | 필수 | “대한민국” “XXX” - don't care |
| 4-3 | BBox[.x[| 바운딩박스 x 좌표 리스트 | number | 필수 | [100, 100, 200, 200] - 4개 |
| 4-4 | BBox[.y[| 바운딩박스 y 좌표 리스트 | number | 필수 | [50, 100, 50, 100] - 4개 |

3 라벨링데이터 실제예시

```

"Annotation": {
  "object_recognition": 0,
  "text_language": 0
},
"Dataset": {
  "category": 0,
  "identifier": "IMG_OCR_53",
  "label_path": "HW_OCR/4.Validation/P.Paper/R.Free/",
  "name": "대용량 손글씨 데이터셋",
  "src_path": "HW_OCR/4.Validation/P.Paper/R.Free/",
  "type": 1
},
"Images": {
  "acquisition_location": "자체",
  "application_field": "기타",
  "background": 0,
  "data_captured": "2021.10.05",
  "height": 3503,
  "identifier": "IMG_OCR_53_4PR_17040",
  "media_type": 0,
  "pen_color": "black",
  "pen_type": 0,
  "type": "png",
  "width": 2475,
  "writer_age": 36,
  "writer_sex": 0,
  "written_content": 1
},
"bbox": [
  {
    "data": "017040",
    "id": 1,
    "x": [1882, 1882, 2179, 2179],
    "y": [139, 220, 139, 220]
  },
  {
    "data": "대구광역시",
    "id": 3,
    "x": [299, 299, 659, 659],
    "y": [1020, 1141, 1020, 1141]
  },
  {
    "data": "저리나드"
  }
]

```

| 데이터셋 구축 수행기관 담당자 | 주관기관 | 기관명 | 책임자명 | 전화번호 (유선전화번호기입) | 메일주소 | 담당업무 |
|------------------------|------|-----------------|----------|--------------------|------|------|
| | | (주)동양시스템즈 | 이태우 | | | |
| | 참여기관 | 기관명 | 담당업무 | 기관명 | 담당업무 | |
| | | (주)쇼우테크 | 수집/가공/검수 | | | |
| | | (주)유니닥스 | 수집/가공/검수 | | | |
| | | 양코르브라보노 협동조합 | 검수 | | | |
| | | | | | | |