

# 데이터 수집

## SKN Family AI Camp 20기 – 최종 프로젝트 3조

프로젝트 주제: 건축 도면 의미 분석 기반 레퍼런스 탐색 및 주소 기반 관련 법규 조회 시스템

---

## 1. 데이터 출처 및 구성

### 1.1 도면 데이터

#### 출처

- AI Hub
- 건축 도면 데이터 → 아파트 → 평면도 데이터

#### 설명

공동주택(아파트) 평면도 이미지와 함께, 공간 및 건축 요소 단위로 라벨링된 학습용 데이터

#### 데이터 형태

- 원천 데이터: PNG 이미지
- 라벨링 데이터: JSON 파일

#### 데이터 분량

- 총 33,998장의 PNG 이미지
  - 각 이미지에 대응되는 라벨링 JSON 데이터 포함
- 

### 1.2 법규 데이터

#### 출처

- 토지이음: 토지이용계획 정보
- 국가법령정보센터

#### 설명

건축 설계 시 필수적으로 검토해야 하는 법적 기준 및 토지 이용 규제 정보

#### 데이터 형태

- JSON 데이터

#### 데이터 수집 방식 및 분량

- 사업지 주소 입력 시마다 실시간 조회되는 정보
  - 웹 크롤링 방식으로 수집
- 

### 1.3 사내 기준 문서 데이터

#### 출처

- 가상의 실사용자 건축사무소를 가정하여 작성한 건축 설계 기준 사내 문서

#### 설명

실제 건축사무소에서 활용될 수 있는 형식의 내부 설계 기준 문서를 가상으로 구성한 데이터

#### 데이터 형태

- .doc / .docx 문서 파일

#### 데이터 분량

- 시험 단계에서는 1~2개 문서를 사용하여 테스트 진행 예정
- 

## 2. 데이터 수집 계기

### 2.1 도면 데이터

건축사무소에서 보관하던 도면을 디지털화하여 관리 효율을 높이고, 도면 내부의 객체 및 공간 구조를 인식하여 자연어 기반 검색이 가능한 조회 방식을 지원함으로써 건축사무소의 실무를 보조하는 기능을 구현하고자 하였다.

국가 AI 개발 지원 플랫폼인 AI Hub에서 제공하는 개방 데이터 중 건축 도면 데이터가 존재함을 확인하였으며, 해당 데이터를 학습한 객체 인식 모델이 함께 제공되고 있음을 확인하였다. 이에 따라 이미지 인식 모델을 처음부터 학습시키는 방식보다, 기존 모델을 활용하고 이를 프로젝트 목적에 맞게 조정(Fine-tuning)하는 방식이 효율적이라고 판단하여 본 데이터를 사용하게 되었다.

---

### 2.2 법규 데이터

해당 사업지의 주소를 기반으로 적용되는 법규를 조회하는 기능을 구현하고자 하였다. 또한 단순 조회에 그치지 않고, 조회된 법규의 실제 적용 여부를 판단하기 위해 법규 체계에 대한 추가적인 검토가 필요하다고 판단하였다.

특히 건축법 체계에서는 상위 법률이 일부 기준을 지방자치단체의 조례에 위임하는 구조가 존재하므로, 이를 고려하기 위해 토지이용, 국가 법령 정보 및 지방자치단체 법규 데이터를 함께 활용하기로 하였다.

---

## 2.3 사내 기준 문서 데이터

본 시스템은 실제 건축사무소 적용을 전제로 하므로, 각 건축사무소가 추구하는 설계 방향성과 기준이 반영된 사내 기준 문서가 필요하다.

이에 따라 사내 기준 문서와 도면 저장 과정에서 추출된 요약 JSON 데이터를 연계하여, 해당 도면이 회사의 설계 방향성과 부합하는지 또는 부적합한지를 평가하는 기능을 구현하고자 하였다.

테스트 및 시연을 위해, 임의의 건축사무소를 가정하고 해당 사무소의 설계 철학과 기준이 반영된 사내 기준 문서를 직접 작성하여 활용할 예정이다.

---

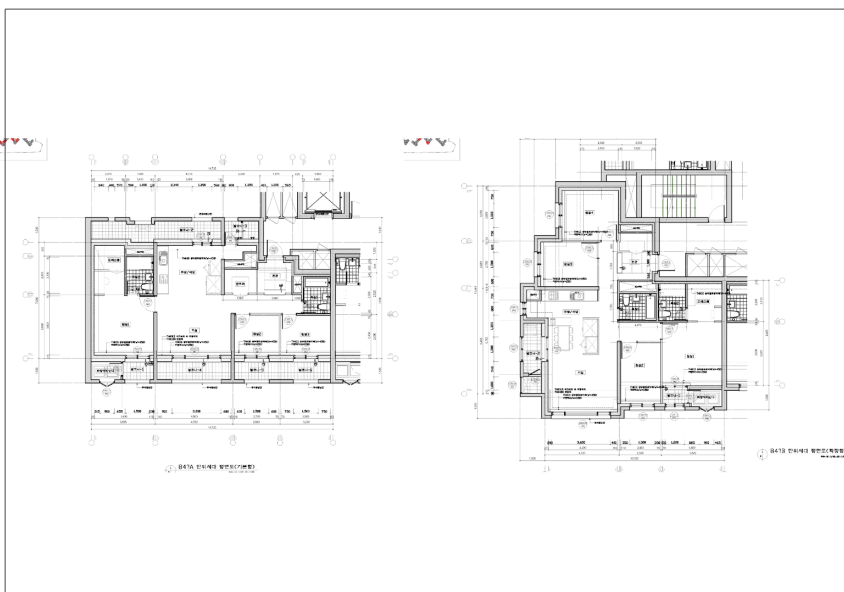
## 3. 데이터 출처 상세

- 도면 데이터  
AI Hub – 건축 도면 데이터 (아파트 평면도)
  - 법규 데이터  
토지이음 – 토지이용계획 정보  
국가법령정보센터
  - 사내 기준 문서 데이터  
건축사무소 사내 문서 작성 시 참고한 논문 및 관련 문서 자료
- 

## 4. 원본 데이터

### 4.1 도면 데이터

- 도면 데이터 (1): 원천 데이터 (PNG 이미지)



- 도면 데이터 (2): 라벨링 데이터 (JSON 파일)

```
{
  "categories": [
    {
      "id": 1,
      "name": "공간_다목적공간"
    },
    {
      "id": 2,
      "name": "공간_엘리베이터홀"
    },
    {
      "id": 3,
      "name": "공간_계단실"
    },
    {
      "id": 4,
      "name": "객체_변기"
    },
    {
      "id": 5,
      "name": "객체_세면대"
    },
    {
      "id": 6,
      "name": "객체_싱크대"
    },
    {
      "id": 7,
      "name": "객체_욕조"
    },
    {
      "id": 8,
      "name": "객체_가스레인지"
    },
    {
      "id": 9,
      "name": "구조_출입문"
    }
  ],
  "images": [
    {
      "id": 54,
      "width": 4961,
      "height": 3508,
      "file_name": "APT_FP_STR_804682526.PNG"
    },
    {
      "id": 55,
      "image_id": 4,
      "category_id": 9,
      "segmentation": [
        [
          2002.1,
          1761.96,
          2116.7,
          1762.06,
          2117.42,
          1631.75,
          2001.5,
          1632.33
        ]
      ],
      "area": 14950.0,
      "bbox": [
        2001.5,
        1631.75,
        115.92,
        130.31
      ],
      "iscrowd": 0,
      "attributes": {
        "구조_출입문": "여닫이문",
        "occluded": false
      }
    }
  ]
}
```

## 4.2 법규 데이터

- 주소 기반으로 조회되는 법규 원본 데이터

```
SKN_3Team_FINAL > land_info_1168010100108250010.json > [ ] 토지이용계획 > [ ] 용도지역_지구
1 {
2   "주소": "서울특별시 강남구 역삼동",
3   "지번": "825-10",
4   "pnu": "1168010100108250010",
5   "공시지가": {
6     "공시지가": 85100000,
7     "기준년도": "2025",
8     "단위": "원/㎡"
9   },
10  "토지이용계획": {
11    "건폐율": "60%",
12    "용적률": "800% / 600% (서울도심)",
13    "용도지역_지구": [
14      "일반상업지역",
15      "용도지역지구"
16    ],
17    "규제법령_기준일": "2025.10.01"
18  },
19  "행위제한": {
20    "높이제한": {
21      "일조권_10m이하_이격거리": "0m",
22      "일조권_10m초과_부분높이": "0"
23    },
24    "규제지역": [
25      "가로구역별_최고높이_제한지역"
26    ],
27    "규제법령": [
28      {
29        "법령명": "건축법",
30        "조항": "제60조",
31        "내용": "건축물의 높이 제한",
32        "전문_전문": "제60조(건축물의 높이 제한) ④ 허가권자는 가로구역(街路區域): 도로로 둘러싸인 일단(一段)의 지역을 말한다. 이하 같다]을 단
33      },
34      {
35        "법령명": "건축법 시행령",
36        "조항": "제82조",
37        "내용": "건축물의 높이 제한",

```

## 5. 데이터 전처리 시 유의 사항

### 5.1 도면 데이터

도면 데이터를 직접 확인한 결과, 동일한 형식의 도면만으로 구성되어 있지 않고 다양한 형태의 도면이 혼재되어 있음을 확인하였다.

- 도면이 잘려 있는 경우가 존재함  
(원본 도면지 경계선에 위치한 문자나 요소가 일부 누락된 사례)
- 하나의 PNG 원천 이미지에 여러 개의 도면이 포함된 경우 존재
- OCR을 통해 문자를 수집할 경우, 도면마다 용어 표기 방식이 상이함

이로 인해 도면 전처리 및 객체·문자 인식 과정에서 추가적인 처리가 필요하다.

---

### 5.2 법규 데이터

법규 데이터 전처리 시, 해당 기준이 어느 법규 계층(법률, 시행령, 시행규칙, 조례 등)에 속하는지 명확히 구분하는 것이 중요하다.

특히 상위 법규에서 특정 기준을 지방자치단체 조례에 위임하고 있는지 여부를 인식하고 구분할 수 있는 구조 설계가 필요하다.

---

### 5.3 사내 기준 문서 데이터

가상의 실사용자 건축사무소 사내 기준 문서를 작성하는 것이므로, 참고 문서는 신뢰성이 높은 자료를 기반으로 선정한다.

문서 작성 시에는 다음 사항을 명확히 기술한다.

- 건축을 통해 거주자의 어떤 점을 고려하고자 하는지
- 해당 목표를 달성하기 위해 어떤 설계 요소를 중점적으로 고려하는지

또한 이러한 기준들이 도면 인식 과정에서 생성된 요약 JSON 정보와 직접적으로 연결될 수 있도록, 기준 항목과 도면 요소 간의 연결 요소를 명확히 정의한다.