

데이터 설명서

- 동의보감 독초판별 이미지 AI 데이터셋 -

담당 역할	기관명	
데이터 구축 총괄	가천대학교 산학협력단	
데이터 설계	(유)엔와이텔	
데이터 수집	(주)믹스미디어, 가천대학교 산학협력단, (주)동북아생물다양성연구소, (유)엔와이텔	
데이터 정제	(유)엔와이텔, 가천대학교 산학협력단	
가공(라벨링, 어노테이션)	(유)나인필스	
데이터 검수	(유)나인필스, (주)아워텍, (유)엔와이텔	
데이터 설명서 작성	(유)엔와이텔	김병철
	가천대학교산학협력단	서정범
데이터 설명서 버전	Ver 1.0 ('21.01.25)	

○ 동의보감 독초판별 이미지 AI 데이터셋

메타데이터 정보 (다중기입가능)	분야	데이터 유형 ¹⁾	구축 데이터량	원천데이터 형식 ²⁾	라벨링 형식 ³⁾	라벨링 유형 ⁴⁾
	<i>N/A 기입</i>	이미지	807,068	jpg	json	바운딩박스(이미지) 폴리곤(이미지)
	데이터 출처 ⁵⁾	데이터 구축년도	구축기관(총 괄)	가공기관	검수기관	
	<i>직접촬영</i>	<i>2021년</i>	가천대학교 산학협력단	나인필스	아이삭, TTA 엔와이텔, 아워텍	
	데이터 문의처	기관명	문의담당자명	전화번호 (유선전화번호기입)	메일주소	
		(유)엔와이텔	김병철	(063) 211-3380	kbcline9@gmail.com	
	데이터 소개	인공지능은 입력된 식물 이미지 데이터만을 가지고 대상의 특징 등을 세분화하고 군집하여 분류하는 과정을 반복적인 훈련을 통해 독초판별 능력을 높이기 때문에, 독초 판별 학습에 필요한 60종의 독초 이미지 데이터셋을 설계하고, 독초와 유사한 식물 61종을 비교 데이터셋으로 추가하여 독초판별의 정확도를 높일 수 있도록 전체 데이터셋을 구성함.				
	주요키워드	동의보감, 독초, 분류 모델, 인공지능, 학습 데이터셋, 알고리즘				
카테고리 정의서		첨부_엑셀				

1) 텍스트, 오디오, 이미지, 비디오,

2) txt, jpg,.....

3) json, csv,.....

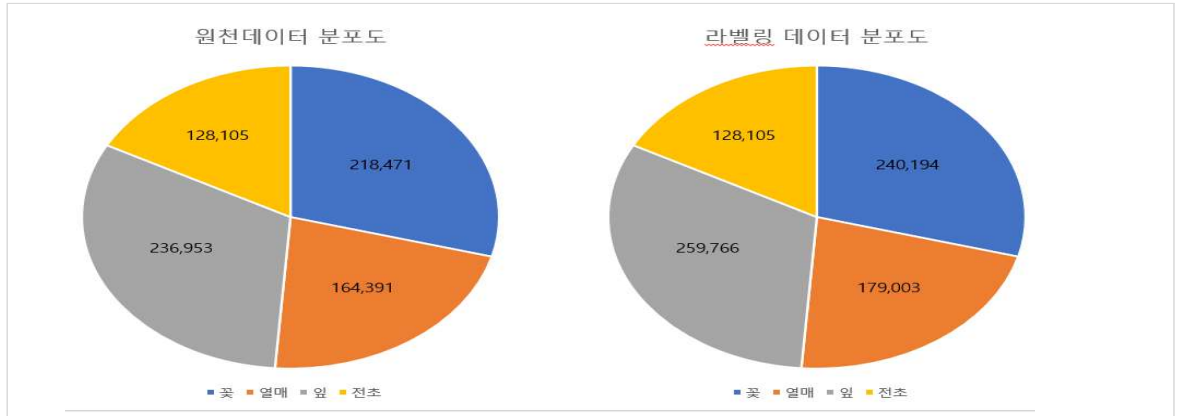
4) 내용요약(텍스트), 번역(자연어), 질의응답(자연어), 바운딩박스(이미지/동영상), 키폰트(이미지/동영상), 세그멘테이션(이미지/동영상), 전자(음성)

5) 4대 언론기사, 자체 수집,.....

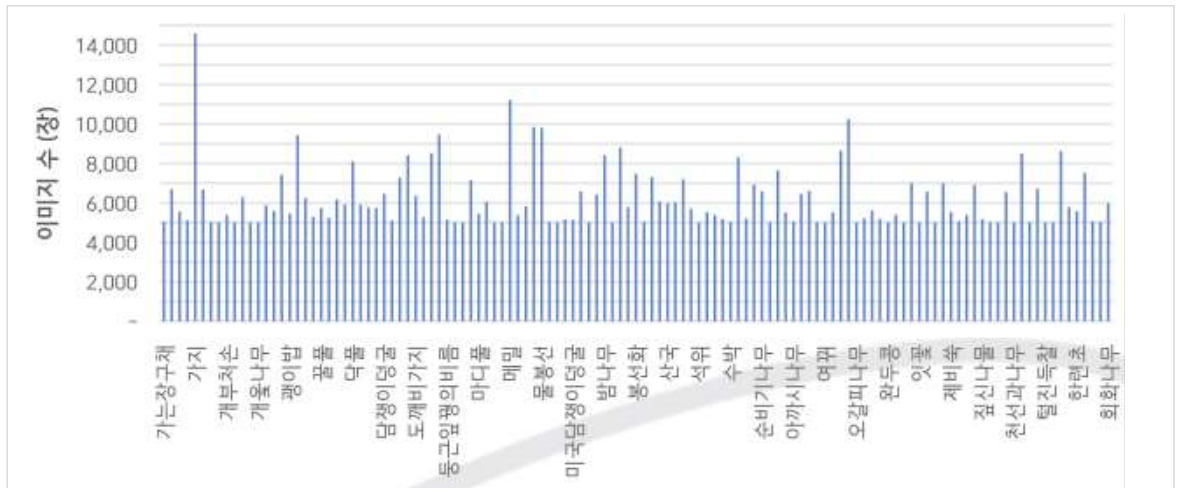
데이터셋명	국립자연사박물관	동의보감 독초판별 이미지 AI 데이터셋																					
		Donguibogam poisonous plants Classification image AI Dataset																					
구축목적	독초에 의한 중독사고 예방을 위해 독초판별이 가능하도록 독초 60종과 유사한 형태를 가진 비교 식물 61종의 전초, 잎, 꽃, 열매 4개 부위에 대해 입체적으로 이미지 데이터를 구축하여 독초판별 인공지능을 훈련하기 위한 데이터셋																						
활용 서비스	인공지능 기술을 활용한 실시간 독초판별 서비스																						
소개	· 사람은 독초나 식물을 판별하기 위해 눈으로 보고(시각), 만져보고(촉각), 냄새 맡고(후각), 맛을 보면서(미각) 종합적이고, 입체적으로 식물의 특징을 파악하고, 자신이 축적된 지식과 경험을 활용하여 종합적으로 판별하며, · 반면 인공지능은 입력된 학습 데이터의 특징과 특성을 분석하고 분류하여 독초나 식물을 판별하기 때문에 이미지 데이터 획득 시 식물의 판별 요소인 잎, 꽃, 열매와 식물의 성장 구분 요소인 전초에 대한 촬영을 입체적으로 할 필요가 있음. · 독초(식물) 판별 방법 - 사람과 인공지능 비교 <table><tr><th>구분</th><th>사람의 경우</th><th>인공지능의 경우</th></tr><tr><td>전초</td><td>전초 생육단계 모양, 형태 등을 관찰하고, 축적된 지식과 판별 경험을 활용</td><td>입력된 전초 이미지 데이터 특성과 특징을 분석</td></tr><tr><td>잎</td><td>잎의 모양, 털의 존재 여부, 잎끝, 잎 밑, 잎 옆, 잎 맥 등을 관찰하고, 축적된 지식과 판별 경험을 활용</td><td>입력된 잎 이미지 데이터 특성과 특징을 분석</td></tr><tr><td>꽃</td><td>차례, 색, 수술 크기, 개화 시기 등을 관찰하고, 축적된 지식과 판별 경험을 활용</td><td>입력된 꽃 이미지 데이터 특성과 특징을 분석</td></tr><tr><td>열매</td><td>열매 색, 모양, 형태, 착과 시기 등을 관찰하고, 축적된 지식과 판별 경험을 활용</td><td>입력된 열매 이미지 데이터 특성과 특징을 분석</td></tr></table> ▼ <table><tr><td>분석 방법</td><td>시각, 촉각, 후각, 미각을 사용하여 입체적으로 관찰하며 분석</td><td>입력된 데이터의 특성과 특징을 분석</td></tr><tr><td>판별 방법</td><td>분석 결과와 축적된 지식, 판별 경험을 종합하여 최종 판별</td><td>분석 결과를 기반으로 학습된 결과와 비교하여 최종 판별 결과 출력</td></tr></table> · 따라서 표에서 보듯이 사람은 촉각, 미각, 후각, 시각을 이용하여 식물을 입체적으로 관찰한 후 자신이 가지고 있는 지식과 경험을 종합하여 독초 여부를 판단하지만, · 인공지능은 입력된 식물 이미지 데이터만을 가지고 판별 대상의 특징을 세분화하고 군집하여 분류하는 과정을 반복적인 훈련을 통해 독초판별 능력을 높이기 때문에, · 한국식물분류학회와 대한 본초학회 전문가 자문을 통해 독초 60종과 독초와 유사한 식물 61종을 비교 데이터셋으로 구성하여 독초판별 정확도를 높일 수 있도록 전체 학습 데이터셋을 설계함.		구분	사람의 경우	인공지능의 경우	전초	전초 생육단계 모양, 형태 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 전초 이미지 데이터 특성과 특징을 분석	잎	잎의 모양, 털의 존재 여부, 잎끝, 잎 밑, 잎 옆, 잎 맥 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 잎 이미지 데이터 특성과 특징을 분석	꽃	차례, 색, 수술 크기, 개화 시기 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 꽃 이미지 데이터 특성과 특징을 분석	열매	열매 색, 모양, 형태, 착과 시기 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 열매 이미지 데이터 특성과 특징을 분석	분석 방법	시각, 촉각, 후각, 미각을 사용하여 입체적으로 관찰하며 분석	입력된 데이터의 특성과 특징을 분석	판별 방법	분석 결과와 축적된 지식, 판별 경험을 종합하여 최종 판별	분석 결과를 기반으로 학습된 결과와 비교하여 최종 판별 결과 출력
	구분	사람의 경우	인공지능의 경우																				
	전초	전초 생육단계 모양, 형태 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 전초 이미지 데이터 특성과 특징을 분석																				
	잎	잎의 모양, 털의 존재 여부, 잎끝, 잎 밑, 잎 옆, 잎 맥 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 잎 이미지 데이터 특성과 특징을 분석																				
	꽃	차례, 색, 수술 크기, 개화 시기 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 꽃 이미지 데이터 특성과 특징을 분석																				
열매	열매 색, 모양, 형태, 착과 시기 등을 관찰하고, 축적된 지식과 판별 경험을 활용	입력된 열매 이미지 데이터 특성과 특징을 분석																					
분석 방법	시각, 촉각, 후각, 미각을 사용하여 입체적으로 관찰하며 분석	입력된 데이터의 특성과 특징을 분석																					
판별 방법	분석 결과와 축적된 지식, 판별 경험을 종합하여 최종 판별	분석 결과를 기반으로 학습된 결과와 비교하여 최종 판별 결과 출력																					
데이터셋 통계 (구축 규모 및 분포)	1. 데이터 구축 규모 <table><tr><th>구분</th><th>세부 내용</th><th>목표</th><th>구축 수량</th></tr><tr><td>수집</td><td>◦ 독초 60종, 부위별 이미지 데이터(jpg) ◦ 유사 비교식물 61종의 부위별 이미지 데이터(jpg)</td><td>600,000장</td><td>1,187,341장</td></tr><tr><td>정제</td><td>◦ 이미지 품질(촬영기준, 정제 메뉴얼) ◦ 독초와 식물이 섞이지 않도록 전문가 전수 검수</td><td>600,000장</td><td>885,531장</td></tr><tr><td>가공</td><td>◦ 공개 저작도구 활용 데이터 가공</td><td>600,000장</td><td>747,869장</td></tr><tr><td>학습 데이터</td><td></td><td>600,000장</td><td>807,068장</td></tr></table>		구분	세부 내용	목표	구축 수량	수집	◦ 독초 60종, 부위별 이미지 데이터(jpg) ◦ 유사 비교식물 61종의 부위별 이미지 데이터(jpg)	600,000장	1,187,341장	정제	◦ 이미지 품질(촬영기준, 정제 메뉴얼) ◦ 독초와 식물이 섞이지 않도록 전문가 전수 검수	600,000장	885,531장	가공	◦ 공개 저작도구 활용 데이터 가공	600,000장	747,869장	학습 데이터		600,000장	807,068장	
구분	세부 내용	목표	구축 수량																				
수집	◦ 독초 60종, 부위별 이미지 데이터(jpg) ◦ 유사 비교식물 61종의 부위별 이미지 데이터(jpg)	600,000장	1,187,341장																				
정제	◦ 이미지 품질(촬영기준, 정제 메뉴얼) ◦ 독초와 식물이 섞이지 않도록 전문가 전수 검수	600,000장	885,531장																				
가공	◦ 공개 저작도구 활용 데이터 가공	600,000장	747,869장																				
학습 데이터		600,000장	807,068장																				

2. 데이터 분포

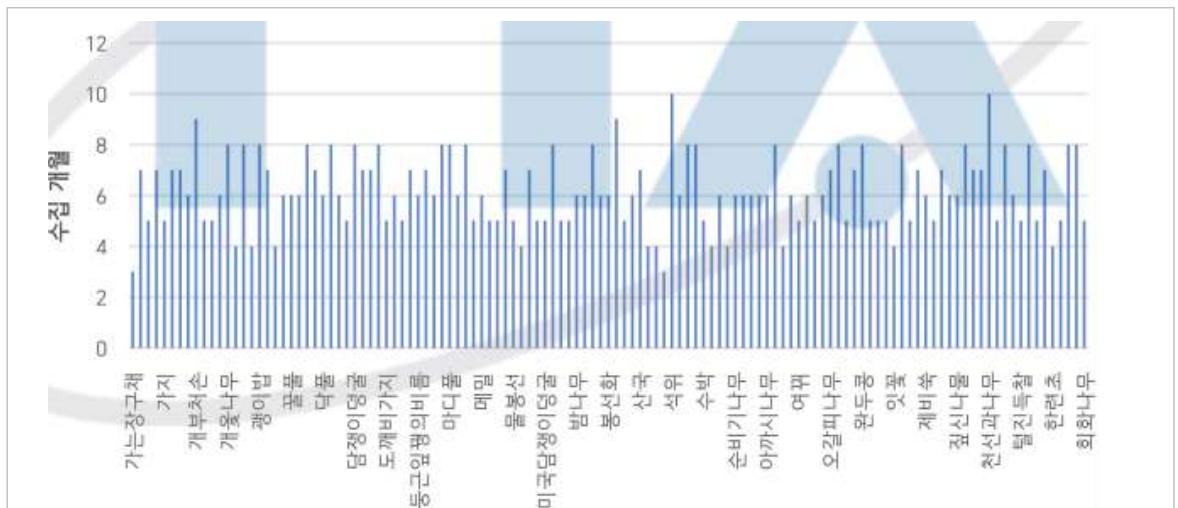
구분	클래스 수	전체 파일 수	인스턴스	인스턴스 수량
원천 데이터	121	747,869	꽃	218,471
			열매	164,391
			잎	236,953
			전초	128,105
라벨링 데이터	121	807,968	꽃	240,194
			열매	179,903
			잎	259,766
			전초	128,105



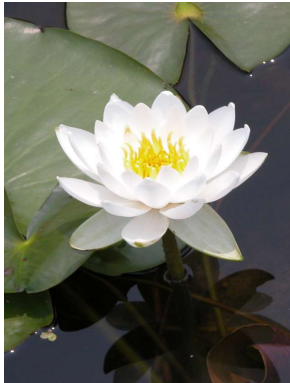
· 클래스 이미지 수량 분포(TTA 최종 품질 검증 결과서 캡처)



· 수집 시기 분포(TTA 최종 품질 검증 결과서 캡처)



1. 대표도면

전초	잎	꽃	열매
			
형식(이미지, jpg)	형식(이미지, jpg)	형식(이미지, jpg)	형식(이미지, jpg)
해상도(HD, FHD)	해상도(HD, FHD)	해상도(HD, FHD)	해상도(HD, FHD)
개체당 24장 이내 촬영	개체당 24장 이내 촬영	개체당 24장 이내 촬영	개체당 24장 이내 촬영

데이터셋	설계	- 독초(60종), 유사 식물(61종) 이미지 속성 정보 - 촬영 도구, 촬영 지역, 촬영 일자 - 컬럼명 표준 기준 적용(외부 자문)	컬럼명 농진청 자문을 통해 표준 컬럼명을 사용함
	수집	- 대상 식물의 부위별 촬영 - 전초, 잎, 꽃, 열매 - 대상 식물의 부위별 압축적 촬영 - 전, 후, 좌, 우(부위별 촬영 기준)	수집은 클라우드워커를 통해 직접 촬영함
	정제	- 이미지 품질 - 분류 오류 감소 - 촬영 기준 감소 - 전문가 전수 감소	정제는 식물분류, 한의학 전문가 의 전수 검수를 수행하 여 최종 정제를 완료함.
	가공	- 공개 저작도구 활용 - Label Me - 바운딩박스 폴리곤(일부 대상)	가공은 클라우드워커를 통해 실 시함
	검수	- 데이터 자체 검수 - 검수 도구 활용(자체 개발) - 외부 전문가 검수 - 주이아이를 통해 품질 검수 수행	검수는 자체 개발한 검수 툴을 이용하여 검수함

(1) 카테고리 정의

- 동의보감 독초판별 이미지 AI 데이터셋 카테고리는 독초 60종과 비교 학습 데이터셋인 유사 식물 61종을 클래스로 하고,
- 각 클래스는 다시 식물을 구성하는 요소인 잎, 꽃, 열매와 식물의 성장 단계를 확인할 수 있는 전초를 인스턴스로 구분하였음.

(2) 클래스 구분

- 클래스는 데이터 공개 시 일반인의 사용성을 높이기 위해 식물명을 우선으로 하고, 인스턴스 또한 데이터 활용 편의성을 보장하기 위하여 한글 명칭을 적용하였음.

클래스 구분	대상 품목	인스턴스	인스턴스 설명
독초 클래스	가죽나무 외 59종	전초	식물의 지상부를 의미함
		잎	식물 판별의 핵심 요소로 식물 성장에 필요한 광합성을 담당하는 기관

비교 클래스	참죽나무 외 60종	꽃	식물의 번식을 담당하는 기관으로 식물 판별의 주요 요소임
		열매	씨방이 발달하여 이루어진 기관으로 식물 판별의 주요 요소임

※ 클래스는 데이터 공개 시 활용성을 높이기 위해 식물명으로 정의하고, 인스턴스 또한 데이터 공개 시 활용 편의성을 보장하기 위하여 한글 명칭을 사용하였음.

· 클래스 및 인스턴스 수

클래스 구분	클래스 수	인스턴스 구분	인스턴스 수	비고
독초 및 유사 비교식물	121	전초, 잎, 꽃, 열매	474	

※ 꽃 부위가 없는 품목 : 무화과, 천선과나무, 넉줄고사리, 황고사리, 석위, 세뿔석위

※ 열매 부위가 없는 품목 : 능소화, 미국 능소화

※ 잎 부위가 없는 품목 : 새삼, 미국실새삼 등은 인스턴스 분류에서 제외됨

2. 라벨링 데이터 구성

속성 구분	속성값	속성값 상세	입력 방안
객체 속성	Class	가죽나무, 풀풀 등 119종	폴더/텍스트 추출입력
	Instance	전초, 잎, 꽃, 열매 등 474	폴더/텍스트 추출입력
이미지 데이터 속성 정보	이미지 ID	가죽나무	텍스트 입력
	파일 유형	JSON	자동 입력
	촬영 지역	시·군·구	텍스트 입력
	촬영 시간	YYYY/MM/DD	추출입력

구분	항목명	타입	필수 여부	설명	범위
1	1	plantinfo	object	클래스 정보	
	1-1	plantinfo.class_name	string	클래스 이름	클래스 이름(121종)
	1-2	plantinfo.classification_info	string	클래스 구분	판별대상, 유사
	1-3	plantinfo.herb_name	string	동의보감 판별대상 이름	동의보감 판별대상 이름
	1-4	plantinfo.similar_plants	list	유사 식물	유사 식물 이름
	1-5	plantinfo.toxic_info	boolean	독성 여부	Y, N
	1-6	plantinfo.instance_info	object	인스턴스 정보	최소 값 : 1, 최대 값 : 1
	1-6-1	plantinfo.instance_info.name	string	식물 부위	전초, 꽃, 잎, 열매
2	2	image	object	인스턴스 세부 정보	이미지 정보, 최소 값 : 7(필드가 Y인 최소 개수
	2-1	image.file_path	string	이미지 경로	../../원천데이터/가죽나무/전초 최소값:1, 최대값 :200
	2-2	image.file_name	string	클래스(식물명)_인스턴스(부위)_file_id	가죽나무_전초_000001.jpg 최소값:4, 최대값:100
	2-3	image.shooting_date	string	촬영 날짜	yyyy-mm-dd

3	2-4	image.region_name	string		촬영 지역	군(시) 또는 장소명 예, (홍천군), (홍릉수목원)
	2-5	image.file_format	string	Y	파일 형식	JPG
	2-6	image.file_size	number	Y	파일 크기	이미지 용량(byte) 최소값:1, 최대값:62914560
	2-7	image.width	number	Y	이미지 너비	Pixel 최소값:1, 최대값:99999
	2-8	image.height	number	Y	이미지 높이	Pixel 최소값:1, 최대값:99999
	3	annotations	array		어노테이션 정보	라벨링 정보
	3-1	annotations[].id	string	Y	라벨링식별자	최소값:1, 최대값:1
	3-2	annotations[].type	string	Y	라벨링 타입	[bbox, polygon]
	3-3	annotations[].bbox	object		바운딩박스 정보	최소값:4, 최대값:4
	3-3-1	bdex_xcrdnt	number		Bounding Box X Coordinate	바운딩박스 X 좌표 최소값:1, 최대값:99999
	3-3-2	bdex_ycrdnt	number		Bounding Box Y Coordinate	바운딩박스 Y 좌표 최소값:1, 최대값:99999
	3-3-3	bdex_width	number		Bounding Box Width	바운딩박스 넓이 최소값:1, 최대값:99999
	3-3-4	bdex_hg	number		Bounding Box Height	바운딩박스 높이 최소값:1, 최대값:99999
	3-4	annotations[].polygon	list		폴리곤 정보	폴리곤 정보
	3-4-1	pyn	number		POLYGON	폴리곤 순번 최소값:1, 최대값:1000
	3-4-2	pyn_xcrdnt	number		POLYGON X Coordinate	폴리곤 X 좌표 최소값:1, 최대값:99999
	3-4-3	pyn_ycrdnt	number		POLYGON Y Coordinate	폴리곤 Y 좌표 최소값:1, 최대값:99999

3 라벨링 데이터 실제 예시

```
{
  "plantinfo": {
    "class_name": "아욱",
    "classification_info": "유사",
    "herb_name": "규채",
    "similar_plants": [
      "난쟁리아욱"
    ],
    "toxic_info": "Y",
    "instance_info": {
      "instance_name": "꽃"
    }
  },
  "image": {
    "file_path": "../..../원천데이터/아욱/전초",
    "file_name": "아욱_전초_887778.jpg",
    "shooting_date": "2021-09-30",
    "region_name": "의성군",
  }
}
```

	<pre> "file_format": "JPG", "file_size": 3548726, "width": 4000, "height": 2252 }, "annotations": [{ "id": "1", "type": "bbox", "bbox": { "bndex_xcrdnt": 335, "bndex_ycrdnt": 700, "bndex_width": 835, "bndex_hg": 500 }, "polygon": null }, { "id": "2", "type": "polygon", "bbox": null, "polygon": [{ "pyn": 1, "pyn_xcrdnt": 590, "pyn_ycrdnt": 1200 }, { "pyn": 2, "pyn_xcrdnt": 1090, "pyn_ycrdnt": 1200 }, { "pyn": 3, "pyn_xcrdnt": 1090, "pyn_ycrdnt": 1600 }, { "pyn": 4, "pyn_xcrdnt": 590, "pyn_ycrdnt": 1600 }] }] </pre>					
--	---	--	--	--	--	--

데이터셋 구축 수행기관 담당자	주관 기관	기관명	책임자명	전화번호 (유선전화번호기입)	메일주소	담당업무
		가천대학교산학협력단	이영종	031-750-5413	garak11@naver.com	사업 총괄
	참여 기관	기관명	담당업무	기관명	담당업무	
		엔와이텔	설계, 데이터 촬영, 정제, 활용	아워텍	AI 알고리즘	
		동북아생물다양성연구소	비교 검색표 작성, 데이터 촬영			
		믹스미디어	데이터 촬영			
		나인펄스	데이터 가공			