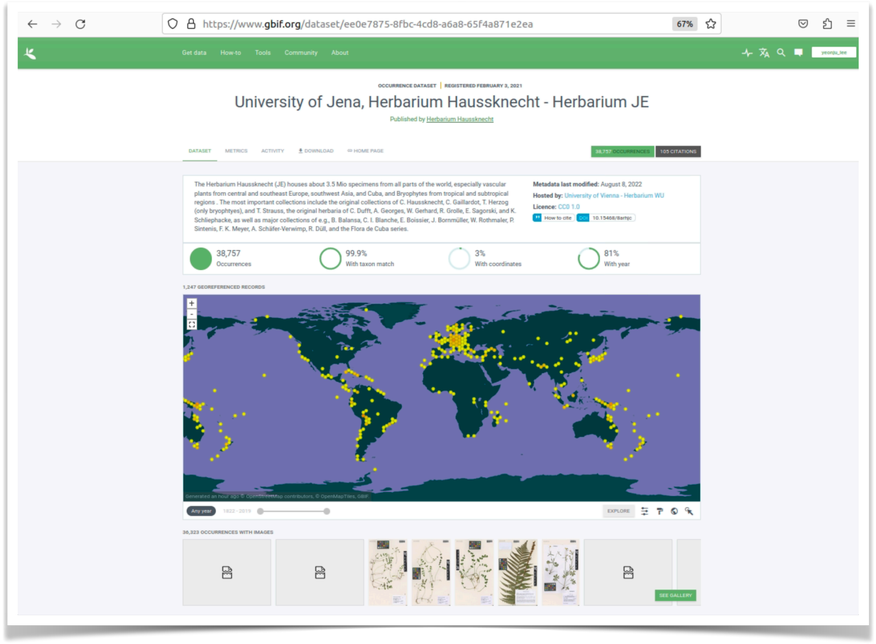
~23.02.28

데이터셋 관련

Haussknecht Herbarium 다운 (30k)

- **Herbarium: 보존된 식물 표본의 모임 (=식물표본집)**
- **저번 시간에 다룬 논문의 데이터셋은 Herbarium으로 일반적인 식물 사진과는 차이가 존재**
- **논문의 데이터셋이 찾음 (https://www.gbif.org/)**



https://www.gbif.org/

~23.02.28

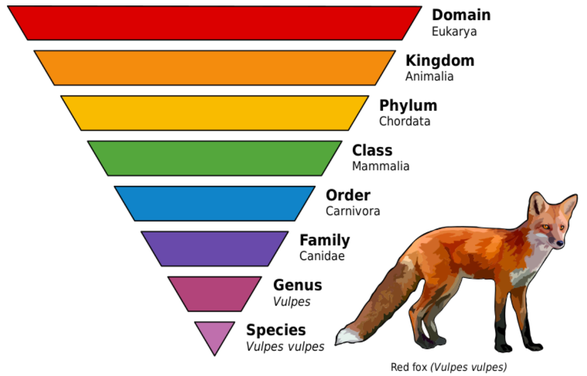
Haussknecht Herbarium

총 50개의 column (gbifID 포함)

- **kingdom ~ species: 종속과목강문계**

  그외에도 date, country 등 다양한 정보가 있음

gbifID	datasetKey	occurrenceID	kingdom	phylum	class	order	family	genus	species
3417438302	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea		Plantae	Tracheophyta	Magnoliopsida	Asterales	Asteraceae	Hieracium	Hieracium lachenalii
3417438301	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea		Plantae	Tracheophyta	Magnoliopsida	Asterales	Asteraceae	Hieracium	Hieracium lachenalii
3414279302	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008834	Plantae	Bryophyta	Bryopsida	Leucodontales	Neckeraceae	Neckera	Neckera intermedia
3414279301	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008835	Plantae	Bryophyta	Bryopsida	Bryales	Roellobryaceae	Roellobryon	Roellobryon roellii
3409209318	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008836	Plantae	Bryophyta	Bryopsida	Bryales	Mniaceae	Mnium	Mnium marginatum
3409209317	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008867	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Cephaloziellaceae	Protolophozia	Protolophozia elongata
3409209316	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008861	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Lophoziaceae	Lophozia	Lophozia savicziae
3409209315	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008871	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Scapaniaceae	Pseudotritomaria	Pseudotritomaria heterophylla
3409209314	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008860	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Lophoziaceae	Lophozia	Lophozia savicziae
3409209313	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008866	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Cephaloziellaceae	Protolophozia	Protolophozia elongata
3409209312	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008864	Plantae	Marchantiophyta	Marchantiopsida	Marchantiales	Aytoniaceae	Plagiochasma	Plagiochasma appendiculatum
3409209311	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008865	Plantae	Marchantiophyta	Marchantiopsida	Marchantiales	Aytoniaceae	Plagiochasma	Plagiochasma appendiculatum
3409209310	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008862	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Lophoziaceae	Lophozia	Lophozia savicziae
3409209309	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008869	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Cephaloziellaceae	Protolophozia	Protolophozia herzogiana
3409209308	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008868	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Cephaloziellaceae	Protolophozia	Protolophozia herzogiana
3409209307	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008859	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Lophoziaceae	Lophozia	Lophozia savicziae
3409209306	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008840	Plantae	Bryophyta	Bryopsida	Bryales	Mniaceae	Mnium	Mnium marginatum
3409209305	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008838	Plantae	Bryophyta	Bryopsida	Bryales	Mniaceae	Mnium	Mnium marginatum
3409209304	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008837	Plantae	Bryophyta	Bryopsida	Bryales	Mniaceae	Mnium	Mnium marginatum
3409209303	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008839	Plantae	Bryophyta	Bryopsida	Bryales	Mniaceae	Mnium	Mnium marginatum
3409209302	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008863	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Lophoziaceae	Lophoziopsis	Lophoziopsis jurensis
3409209301	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE04008870	Plantae	Marchantiophyta	Jungermanniopsida	Jungermanniales	Cephaloziellaceae	Protolophozia	Protolophozia herzogiana
3403694346	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE00034432	Plantae	Tracheophyta	Magnoliopsida	Asterales	Asteraceae	Baccharis	Baccharis macraei
3403694345	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE00034709	Plantae	Tracheophyta	Magnoliopsida	Asterales	Asteraceae		
3403694344	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE00034717	Plantae	Tracheophyta	Magnoliopsida	Asterales	Asteraceae		
3403694343	ee0e7875-8fbc-4cd8-a6a8-65f4a871e2ea	https://je.jacq.org/JE00034713	Plantae	Tracheophyta	Magnoliopsida	Asterales	Asteraceae		

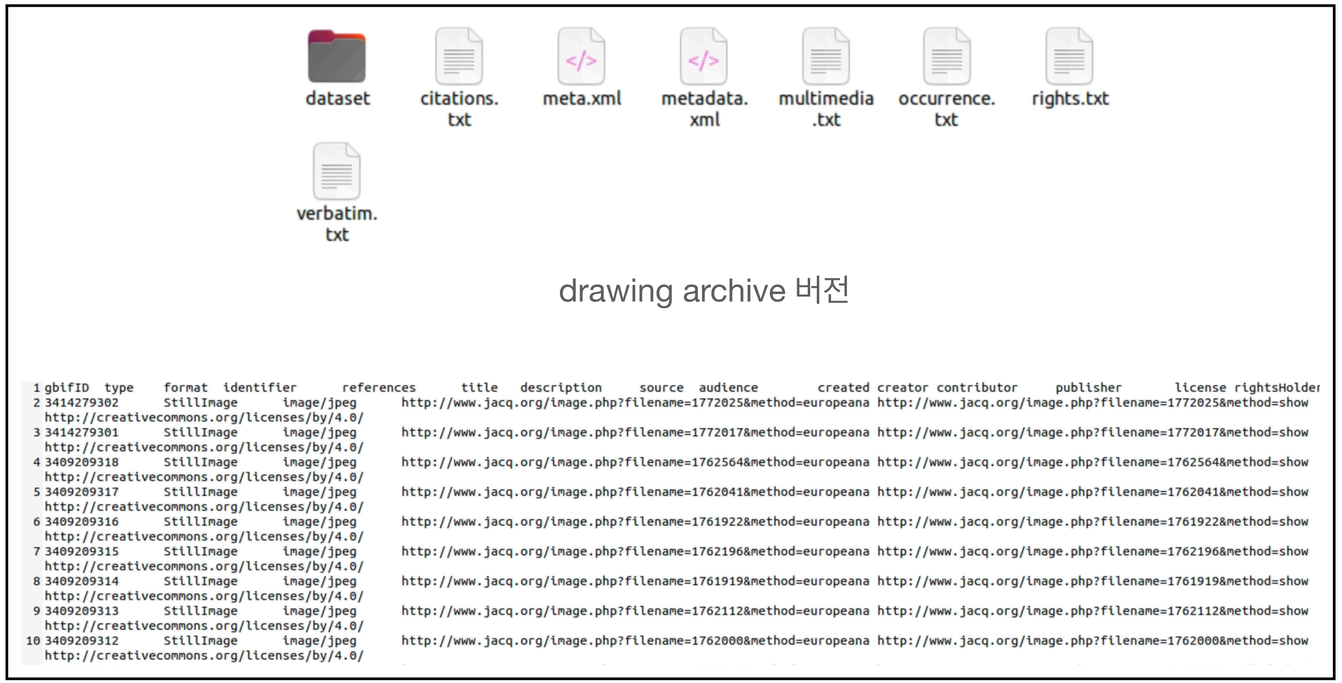


두 가지 버전

simple & darwin archive version

- 이미지를 따로 구하기 쉽지 않음 - 크롤링이나 외부 플러그인 설치 필요 (이미지 사용 시, 따로 동의 필요)

simple 버전



drawing archive 버전

	gbifID	type	format	identifier	references	title	description	source	audience	created	creator	contributor	publisher	license	rightsHolder
2	3414279302	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1772025&method=europeana	http://www.jacq.org/image.php?filename=1772025&method=show	http://creativecommons.org/licenses/by/4.0/									
3	3414279301	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1772017&method=europeana	http://www.jacq.org/image.php?filename=1772017&method=show	http://creativecommons.org/licenses/by/4.0/									
4	3409209318	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1762564&method=europeana	http://www.jacq.org/image.php?filename=1762564&method=show	http://creativecommons.org/licenses/by/4.0/									
5	3409209317	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1762041&method=europeana	http://www.jacq.org/image.php?filename=1762041&method=show	http://creativecommons.org/licenses/by/4.0/									
6	3409209316	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1761922&method=europeana	http://www.jacq.org/image.php?filename=1761922&method=show	http://creativecommons.org/licenses/by/4.0/									
7	3409209315	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1762196&method=europeana	http://www.jacq.org/image.php?filename=1762196&method=show	http://creativecommons.org/licenses/by/4.0/									
8	3409209314	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1761919&method=europeana	http://www.jacq.org/image.php?filename=1761919&method=show	http://creativecommons.org/licenses/by/4.0/									
9	3409209313	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1762112&method=europeana	http://www.jacq.org/image.php?filename=1762112&method=show	http://creativecommons.org/licenses/by/4.0/									
10	3409209312	StillImage	image/jpeg	http://www.jacq.org/image.php?filename=1762000&method=europeana	http://www.jacq.org/image.php?filename=1762000&method=show	http://creativecommons.org/licenses/by/4.0/									

multimedia.txt (image 주소)

Herbarium Dataset

문제(?)점

- **Herbarium Dataset에서 bounding box가 된 채 배포 x**

researchers and the public. The collected data contain objects with a high degree of variability in scale and occlusion, making it one of the most challenging data sets. Among them, we annotated manually 4000 images having specimen images with distorted leaves (leaves with missed part) or overlapping leaves, not only specimens containing perfect leaves and some samples are shown in Figure 5. Herbarium specimen images contain seven main regions of scale-bar, barcode, stamp, annotation label, color pallet, envelope, and the plant specimen. Every region is represented by a bounding box described by x, y, width and height within the XML file (Figure 3). We emphasize that the bounding box is dedicated by annotating all objects within the digitized specimen images except the plant specimen region. Otherwise, because of its irregular shape, we describe the plant specimen region by a bounding polygon.

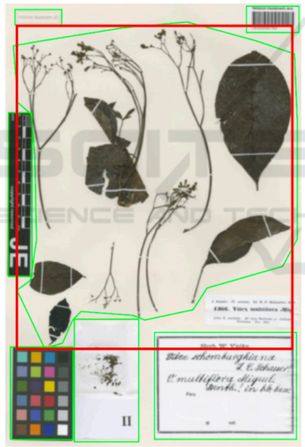


Figure 3: Annotation process of a digitized specimen image.

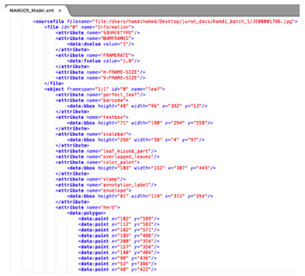


Figure 4: Example of the XML annotation file.

Figure 5: Examples of herbarium Haussknecht dataset.

The dataset was divided into 80% training set, 10% validation set and 10% test set. We trained the original YOLO V3 and improved YOLO V3 models on DHS database. In both networks, the parameters are set as follows: the initial learning rate is reduced to 0.0001 and batch size is 6. Furthermore, all networks were trained for 10000 iterations and we got the avg loss curve as presented in Figure 7.

4 RESULTS AND EVALUATION

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

PMC PubMed Central® Search PMC Full-Text Archive Search in PM

Run this search in PubM

Journal List > Biodivers Data J > v.8; 2020 > PMC7746675

Pensoft Publishers | Biodiversity Data Journal | About | Editorial Team | Author Guidelines | Submission

Biodivers Data J. 2020; 8: e57090. PMCID: PMC7746675
Published online 2020 Dec 10. doi: 10.3897/BDJ.8.e57090 PMID: 33343217

Detection and annotation of plant organs from digitised herbarium scans using deep learning

Sohaib Younis,[✉1,2] Marco Schmidt,[3,1] Claus Weiland,[1] Stefan Dressler,[4] Bernhard Seeger,[2] and Thomas Hickler[1]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

Associated Data

▸ Supplementary Materials

Abstract Go to: ▸

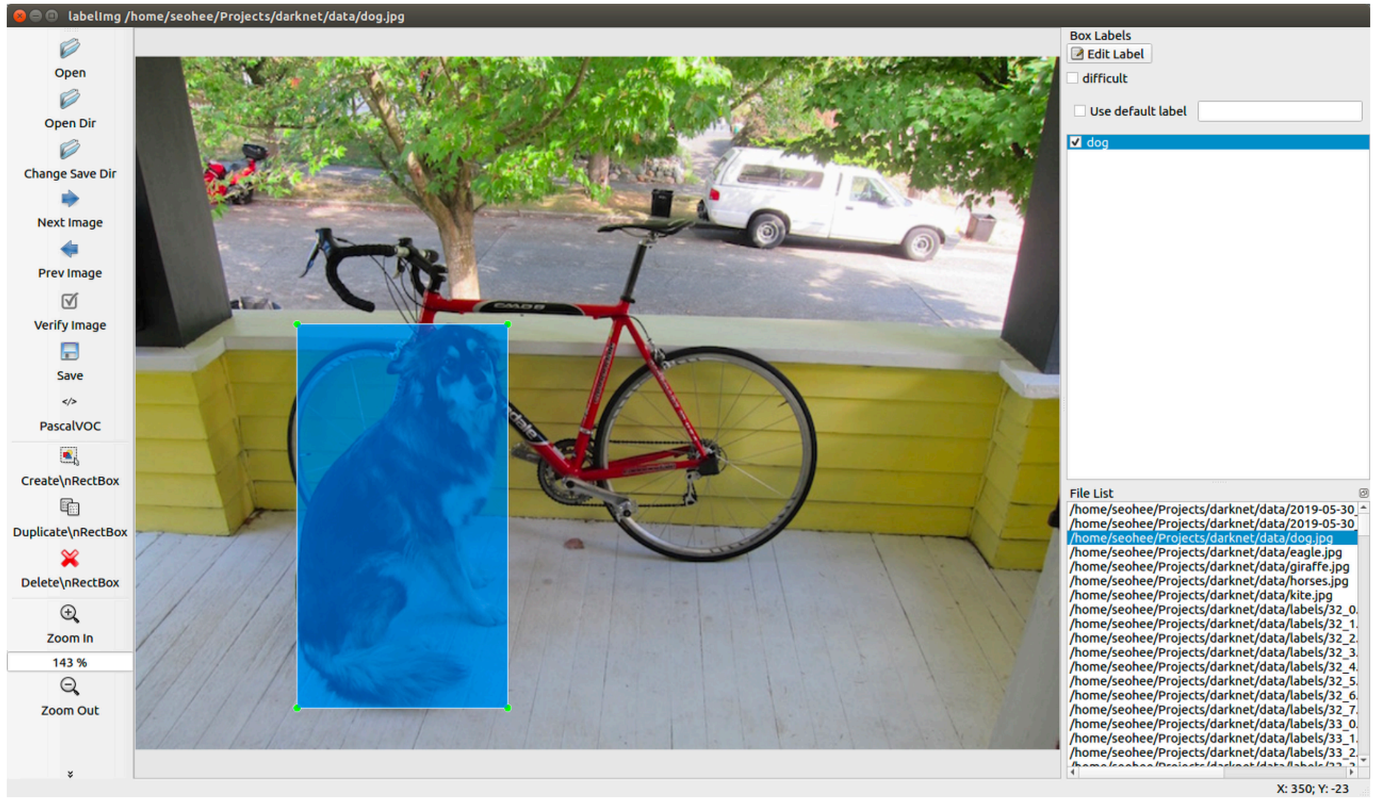
As herbarium specimens are increasingly becoming digitised and accessible in online repositories, advanced computer vision techniques are being used to extract information from them. The presence of certain plant organs on herbarium sheets is useful information in various scientific contexts and automatic recognition of these organs will help mobilise such information. In our study, we use deep learning to detect plant organs on digitised herbarium specimens with Faster R-CNN. For our experiment, we manually annotated hundreds of herbarium scans with thousands of bounding boxes for six types of plant organs and used them for training and evaluating the plant organ detection model. The model worked particularly well on leaves and stems, while flowers were also present in large numbers in the sheets, but were not equally well recognised.

Biodivers Data J

manual image annotation tools

example

- **https://github.com/tzutalin/labelImg**



dataset 조사 (1)

종분류 o, bbox x

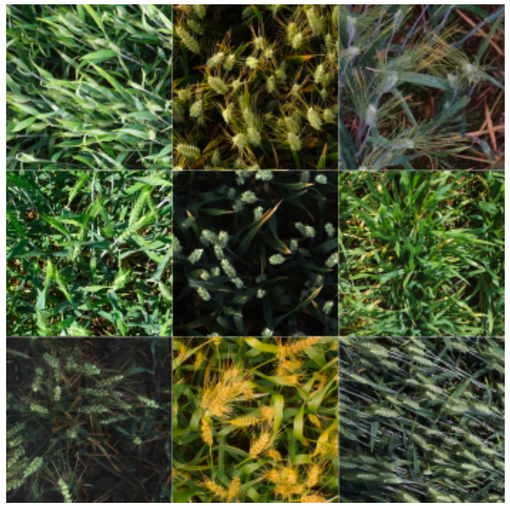
- Herbarium 2021 - Half-Earth Challenge - FGVC8
  - JPEG
  - JSON files in COCO dataset format
  - 161.9 GB
  - 처음 소개된 Walker 논문의 dataset
- Herbarium 2022 - FGVG9

```
annotation {
  "image_id" : int,
  "category_id" : int,
  "genus_id" : int,
  "institution_id" : int
}

image {
  "image_id" : int,
  "file_name" : str,
  "license" : int
}

category {
  "category_id" : int,
  "scientificName" : str,
  # We also provide a super-category for each species.
  "authors" : str, # correspond to 'authors' field in the wcvp
  "family" : str, # correspond to 'family' field in the wcvp
  "genus" : str, # correspond to 'genus' field in the wcvp
  "species" : str, # correspond to 'species' field in the wcvp
}

genera {
  "genus_id" : int,
  "genus" : str
}

distance {
  # We provide the pairwise evolutionary distance between categories (genus_id0 < genus_id1).
  "genus_id_x" : int,
  "genus_id_y" : int,
  "distance" : float
}
```

dataset 조사 (2)

종분류 x, bbox o, herbarium x -> classes = 0으로 진행 가능

- **Global Wheat Detection - kaggle**



Columns

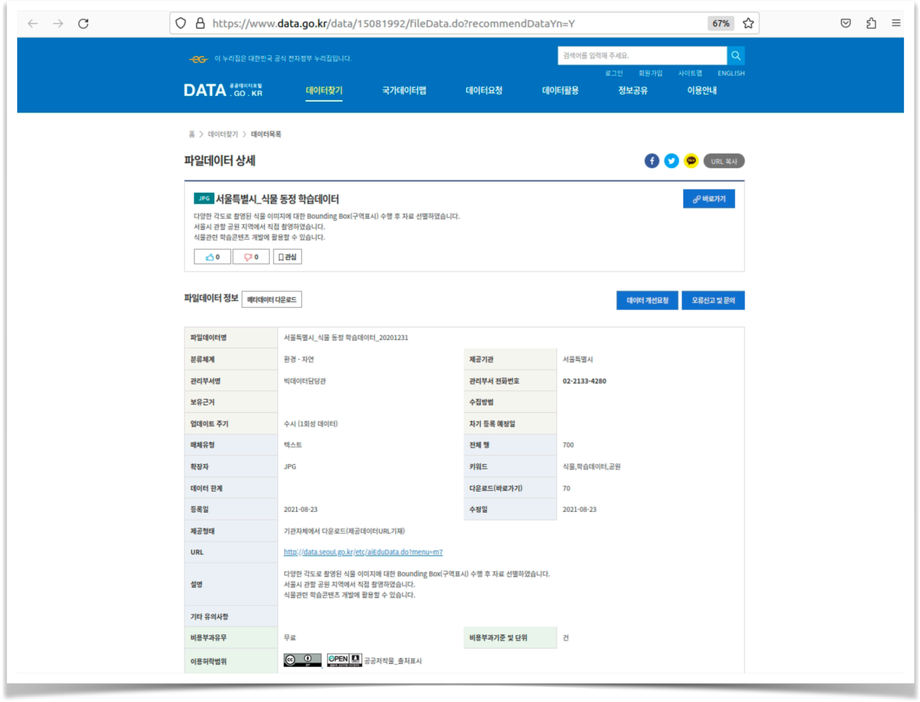
- `image_id` - the unique image ID
- `width`, `height` - the width and height of the images
- `bbox` - a bounding box, formatted as a Python-style list of [xmin, ymin, width, height]
- etc.

  - **643.57 MB**
  - **jpg, csv**

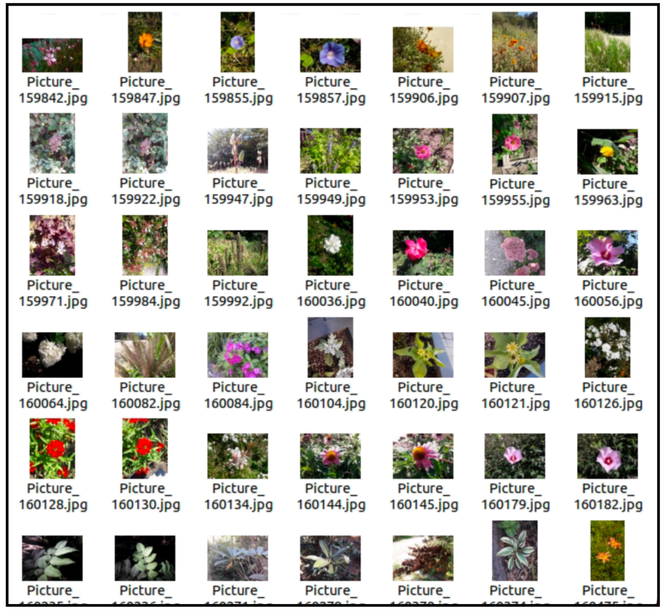
dataset 조사 (3)

종분류 x, bbox o, herbarium x -> classes = 0으로 진행 가능

- 서울특별시_식물 동정 학습데이터
  - 1.4 GB
  - jpg, xml



https://www.data.go.kr



image

```xml
<annotation>
	<folder>20200921 ŽçÇöÃµ</folder>
	<filename>Picture_159842.jpg</filename>
	<path>C:\Users\infoboss\Downloads\20200921 ŽçÇöÃµ\Picture_159842.jpg</path>
	<source>
		<database>Unknown</database>
	</source>
	<size>
		<width>4160</width>
		<height>2340</height>
		<depth>3</depth>
	</size>
	<segmented>0</segmented>
	<object>
		<name>flower</name>
		<pose>Unspecified</pose>
		<truncated>0</truncated>
		<difficult>0</difficult>
		<bndbox>
			<xmin>1318</xmin>
			<ymin>906</ymin>
			<xmax>2439</xmax>
			<ymax>2118</ymax>
		</bndbox>
	</object>
</annotation>
```

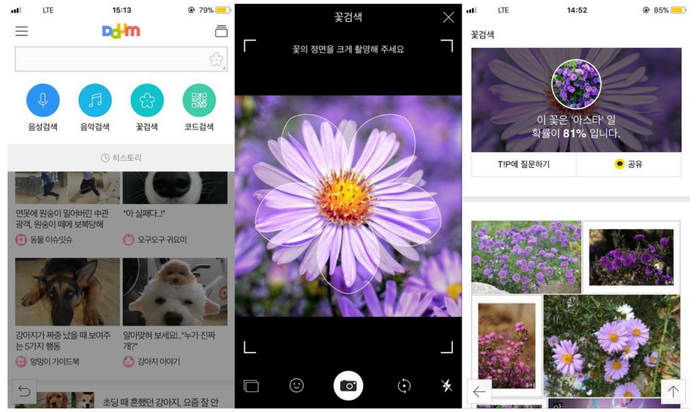
각 image의 xml 파일

다음 / 네이버 꽃 검색

다음 꽃 검색 - AI (현재도 서비스 중)

카카오AI리포트 (2018)

https://brunch.co.kr/@kakao-it/260



[그림 1] 다음 앱 꽃검색 서비스 화면*1

- 2015, 국내에 주로 서식하는 약 500여 가지의 꽃 품종에 대한 꽃 사진 수집을 통해 십여 만장의 꽃 품종 분류 데이터 셋(data set)을 구축
- 딥러닝 이전에도 전통적인 기계학습 (SVM)을 통한 시도 o
- 품종별 사진 데이터 수의 불균형 문제(data imbalance problem)
  - 사진의 수가 적은 품종을 기준으로 다른 품종의 사진을 무작위로 제거하여 수를 줄여주거나(undersampling), 사진의 수가 많은 품종을 기준으로 다른 품종의 사진을 변형해 수를 늘려주는 방법(oversampling)
- 최종적으로 구축된 10만 장 이상의 대규모 꽃 사진 데이터 셋은 2015년 당시 가장 높은 성능을 보인 CNN 모델을 바탕으로 한 꽃 품종 분류 네트워크를 학습하기 위해 사용
  - 꽃검색 서비스에 꽃이 아닌 사진이 제공되는 경우 임의의 꽃 품종을 출력하는 경우를 막기 위해 사진의 꽃 여부를 판단하는 이진 분류기를 추가로 학습 및 적용
  - 다양한 데이터 증대 기법을 이용해 꽃 사진 데이터 셋을 변형함으로써 각도, 날씨, 조명 등 보다 다양한 꽃 사진 촬영 환경에 대해 강인한 분류 성능을 얻을 수 있었던 것
- 기존 전통적인 컴퓨터 비전 기술을 이용한 꽃 검색 알고리즘: 102종 분류도 충분한 정확도 x
- 딥러닝을 이용한 꽃 품종 분류 알고리즘: 500종이 넘는 품종을 분류, 기존 대비 분류 정확도를 30% 이상 개선