



Data Gathering & Annotation

Andrea Filibeto Lucas

Quick Introduction...



What is Data Gathering?

The process of **collecting raw information** from various sources e.g., cameras, sensors or online datasets)

In AI projects, this step provides the **foundation for model training.**

What is Data Annotation?

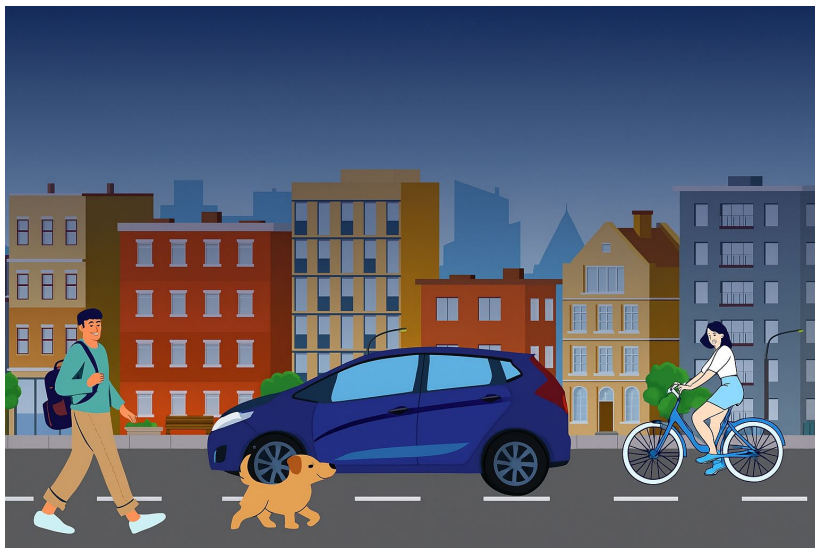
The process of **labelling collected data to give it meaning and structure** for machine learning.

Each data sample is **tagged with relevant information** (e.g., bounding boxes, categories, or attributes).

Or More Simply...

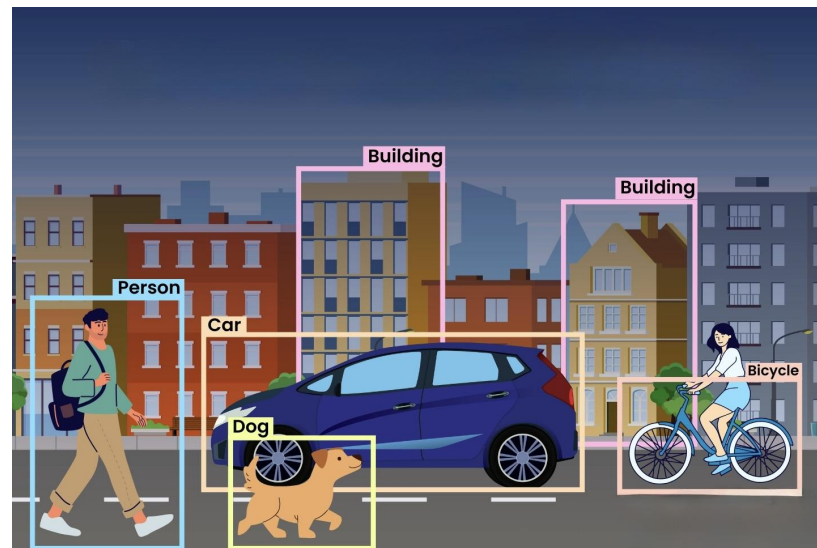


Data Gathering



Sourced from: www.projectpro.io/article/data-annotation-in-ai/1154

Data Annotation



NB: This is an incomplete annotation - not all buildings/people are annotated.

Types of Annotation



1. Image Description

Provides a textual summary of the image.

Common in captioning datasets (e.g., “A red stop sign on a rural road”).

2. Box Annotation

Uses bounding boxes to mark specific objects.

Essential for object detection tasks (e.g., locating traffic signs).

3. Segmentation

Divides the image into pixel-level regions.

Used for precise object boundaries (e.g., segmenting a road sign’s exact shape).

4. Keypoint & Landmark Annotation

Marks specific points of interest (e.g., facial features, joints, or pole mounting points).

Good vs Bad Annotation



Bounding boxes **TIGHTLY FIT** around objects.

Correct labels are applied.

Consistent style across images (same criteria, naming, and labelling precision).

All visible instances are annotated - none are missed.

Why This Matters?

Poor annotation leads to misleading training data, causing models to learn incorrect patterns and perform unreliably in real-world conditions.



Annotation platforms

What are Annotation Platforms?



L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence

Annotation platforms are **software tools** designed to label and manage datasets for ML.

They allow users to **draw boxes**, masks, or keypoints on images, assign class labels, and **export data in standard formats**.

These platforms streamline the process of preparing high-quality, structured data for training CV models.

Examples:

- [Label Studio](#), CVAT (CV Annotation Tool) or Roboflow



NB: For your assignment, you must use Label Studio.

Export Annotation Formats?



COCO Format

- Stores *images*, *annotations*, and *categories* in one file.
- Used by YOLOv8, TensorFlow, etc.

YOLO Format

- One **.txt** per image: **class x_center y_center width height**.
- Compact, normalised (0–1), ideal for real-time detection.

Label Studio / JSON

- Keeps full metadata and attributes.

Conversion

- Libraries like **Ultralytics** or **custom scripts** convert between formats.
- **Always check coordinate systems after conversion.**

```
1  {
2      "id": 3,
3      "image_id": 2,
4      "category_id": 1,
5      "bbox": [
6          124.06,
7          86.66,
8          90.61,
9          85.87
10     ],
11     "area": 7780.68,
12     "iscrowd": 0,
13     "attributes": {
14         "view_angle": "Front",
15         "mounting": "Pole-mounted",
16         "condition": "Good",
17         "sign_shape": "Circular"
18     }
19 }
```

(COCO Annotation Sample)



Gathering Good Data

Good vs Bad Data Gathering



L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence



Front



Side



Back

Good vs Bad Data Gathering



L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence



Front



Side



Back



Case: Sign not installed correctly



L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence



Front



Side



Back

Case: Sign not installed correctly



L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence



Front



Side



Back

Case: Sign not installed correctly



L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence



Front



Side



Back



Poor data gathering, such as capturing visible number plates or unclear sign views, violates GDPR standards and reduces dataset quality. Signs should be clearly visible (even if not centred).



Setting-Up Label Studio

Please follow the demonstration.

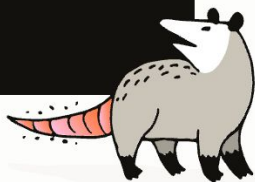
Getting Started: Installation



Quick Start

PIP BREW GIT DOCKER

```
1 # Install the package
  # into python virtual environment
2 pip install -U label-studio
3
4 # Launch it!
5 label-studio
```



Quick Start

PIP BREW GIT DOCKER

```
1 # Install the cask
2 brew install humansignal/tap/label-studio
3
4 # Launch it!
5 label-studio
```



Ensure you have Python (up to version 3.12) & pip/brew installed on your system.

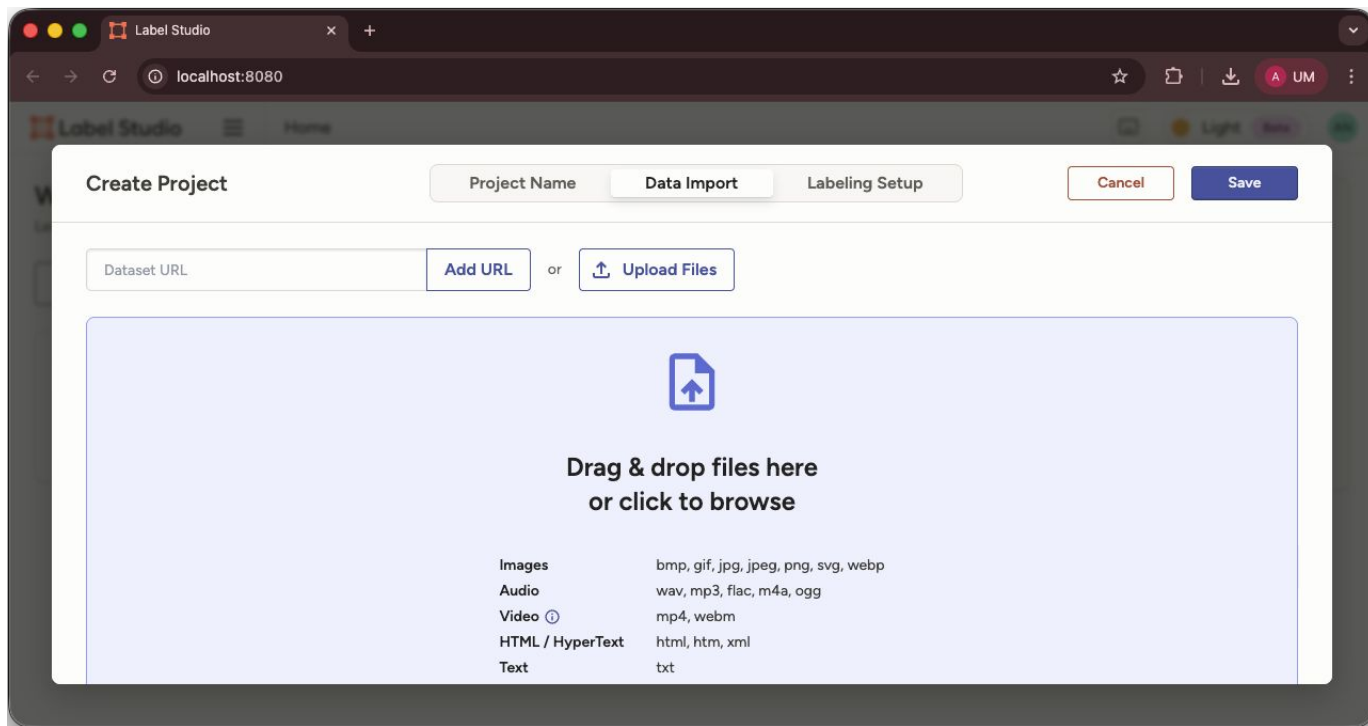
For more information visit: <https://labelstud.io/>

Step 1: Create a New Project

A screenshot of a web browser window showing the Label Studio application. The browser's address bar displays 'localhost:8080'. The application's 'Create Project' modal is open, featuring three tabs: 'Project Name', 'Data Import', and 'Labeling Setup'. The 'Project Name' tab is active, showing a text input field with 'New Project #1', a larger text area for a description with the placeholder 'Optional description of your project', and a 'Workspace' dropdown menu currently set to 'Enterprise'. At the bottom of the modal, there is a link to 'Learn more' about project management. 'Cancel' and 'Save' buttons are located in the top right corner of the modal.

After starting Label Studio (at <http://localhost:8080> - by default), click "Create Project".

Step 2: Import Your Data



You can upload files (images, text, CSVs) directly from your computer, or you can connect to cloud storage.

Step 3: Set Up Labeling Interface



Label Studio

localhost:8080/projects/2/settings/labeling

Projects / ASN_YR3_25/26 / Settings / Labeling Interface

Labeling Interface

Browse Templates

Code Visual

```
1 <View>
2   <Header value="Maltese Traffic Signs: Detection and Att
3   |
4   <!-- OBJECT -->
5   <View style="display:flex;align-items:start;gap:8px;flex:
6   <Image name="image" value="$image" zoom="true" zoomCon
7   <RectangleLabels name="sign_type" toName="image" show:
8     <Label value="Stop" category="1" background="#E53939"
9     <Label value="No Entry (One Way)" category="2" backg
10    <Label value="Pedestrian Crossing" category="3" back
11    <Label value="Roundabout Ahead" category="4" backgro
12    <Label value="No Through Road (T-Sign)" category="5"
13    <Label value="Blind-Spot Mirror (Convex)" category="
14  </RectangleLabels>
15 </View>
16
```

Configure the labeling interface with tags.
See all available tags.

Save

Additional Tags

Maltese Traffic Signs: Detection and Attribute Labelling

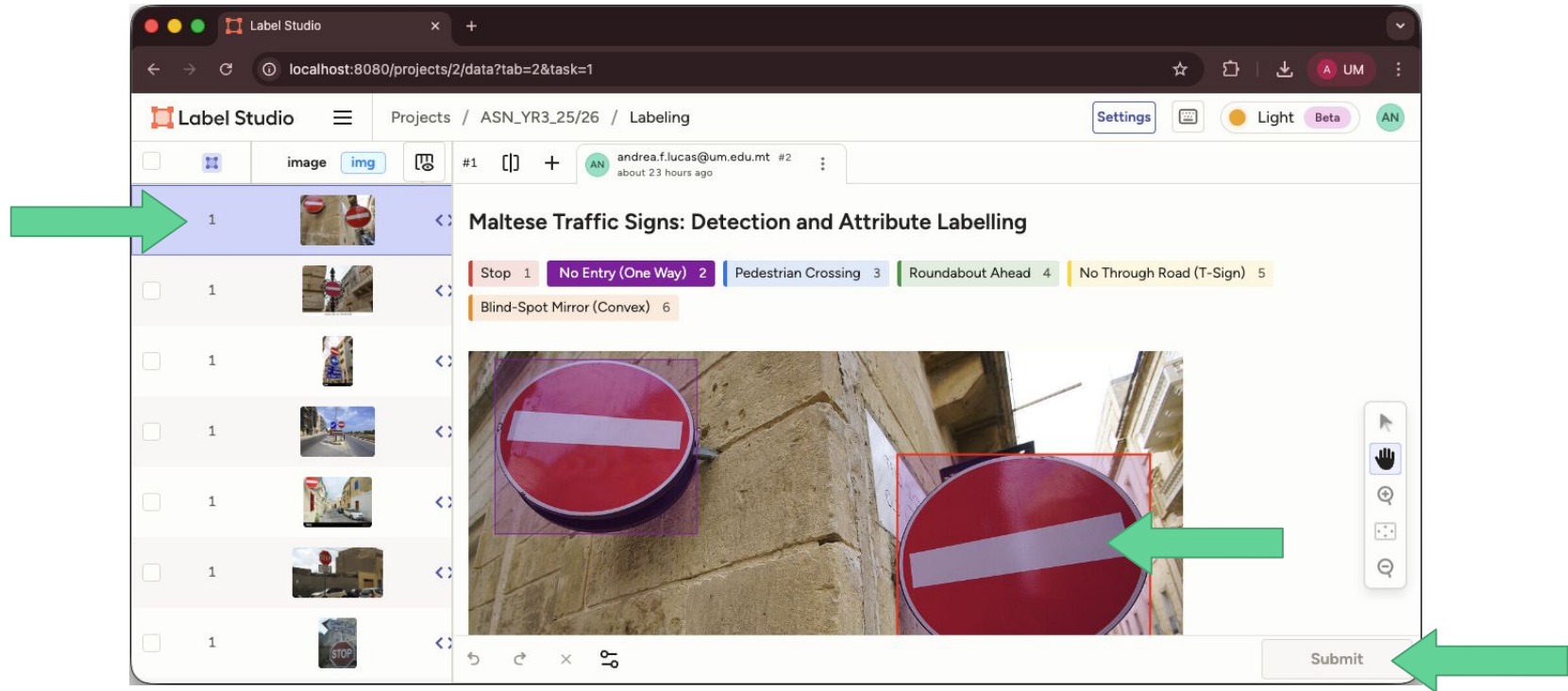
Stop 1 No Entry (One Way) 2 Pedestrian Crossing 3

Roundabout Ahead 4 No Through Road (T-Sign) 5

Blind-Spot Mirror (Convex) 6

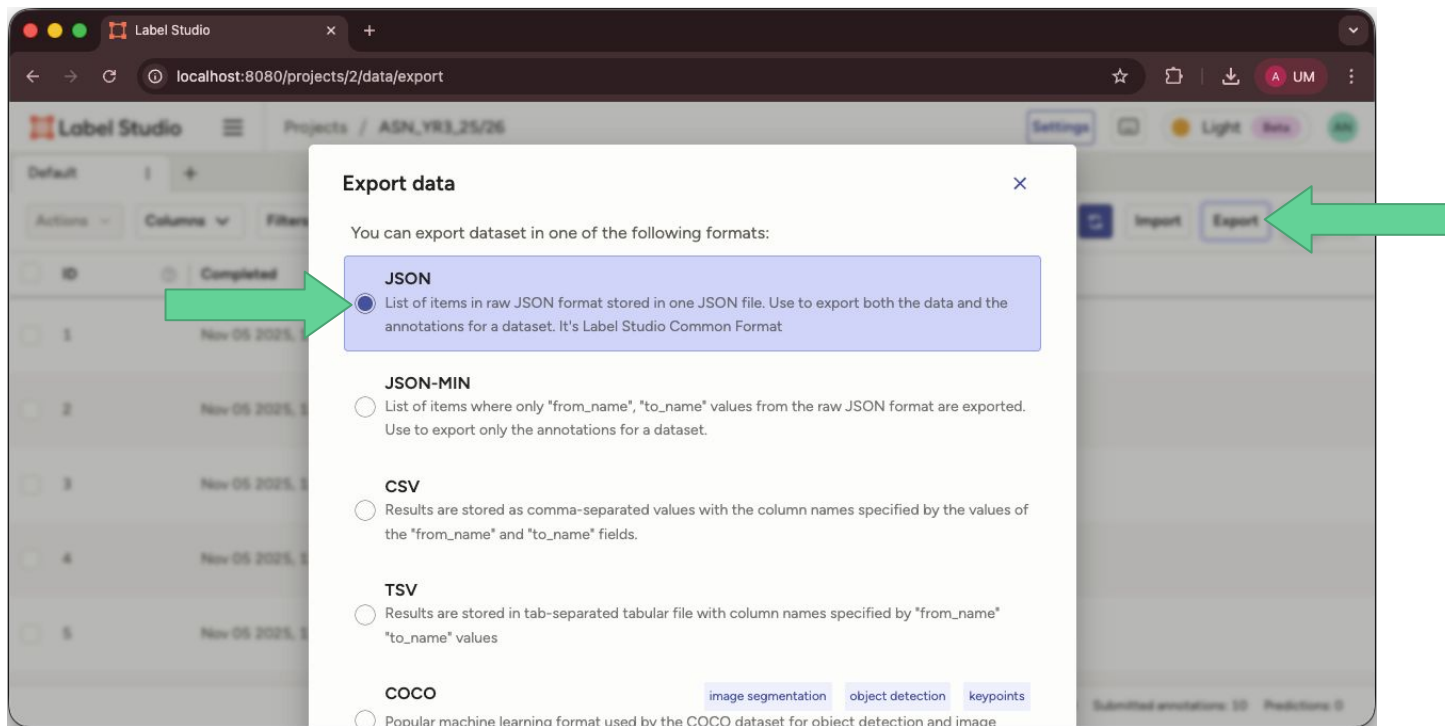
For your assignment, the XML file will be provided to you.

Step 4: Start Labeling



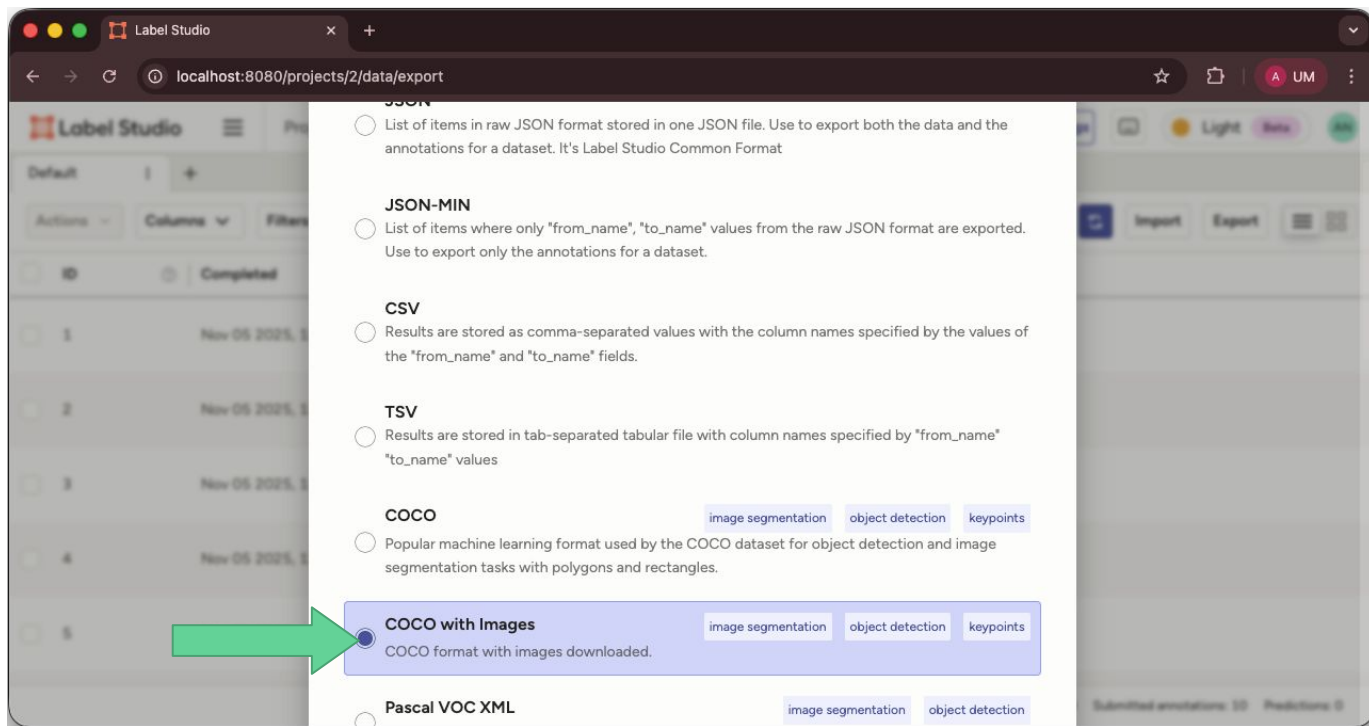
Click "Label All Tasks" from the project dashboard begin. Select your labels (e.g., click "No Entry") and press "Submit".

Step 5a: Export Your Annotated Data



Once you've labeled your data, click the "Export" button on the project page & choose your desired format.

Step 5b: Export Your Annotated Data



Scroll down and export using 'COCO with Images' to get the updated image filenames.



Important for your assignment: you need JUST the images folder from the COCO (with images) export and the JSON file from the Label Studio JSON export. Please follow the provided notebook carefully for the conversion to work correctly.

Step 6: Convert Annotations



Input



JSON Export



*Images Folder as .zip
from COCO with Images*

Conversion Script



Output



COCO.json



[sign_attribute].json

NB: Use the provided Jupyter notebook with the JSON annotations and corresponding images (as a .zip file) to generate the assignment's COCO-format annotations.



Any Questions?



Andrea Filiberto Lucas



andrea.f.lucas@um.edu.mt



Dylan Seychell



dylan.seychell@um.edu.mt

Further Reading & Resources



- <https://labelstud.io/>
- <https://roboflow.com/>
- <https://www.cvat.ai/>
- <https://www.youtube.com/watch?v=R1ozTMrujOE> *(Label Studio Tutorial)*
- <https://blog.roboflow.com/how-to-use-label-studio/>
- <https://labelstud.io/learn/getting-started-with-label-studio/>
- <https://blog.roboflow.com/yolov11-how-to-train-custom-data/>