

# WebLI

**Data Card Authors:** Xiao Wang , Keran Rong

## DATASET SUMMARY

**WebLI** (Web Language Image) is a web-scale multilingual image-text dataset, designed to support Google’s vision-language research, such as the large-scale pre-training for image understanding, image captioning, visual question answering, object detection etc.

The dataset is built from the public web, including image bytes, image-associated texts (alt-text, OCR, page title), 109 languages and many other features. The dataset is deduplicated on 68 common vision/vision-language tasks, and has no user or personally identifiable data with careful RAI considerations.

## Dataset Overview

DATA SUBJECT(S)	DATASET SNAPSHOT	CONTENT DESCRIPTION																
<div>Sensitive Data about people</div> <div>Non-Sensitive Data about people</div> <div>Data about natural phenomena</div> <div>Data about places and objects</div> <div>Synthetically generated data</div> <div>Data about systems or products and their behaviors</div> <div>Unknown</div> <div>Others (Please Specify)</div>	<table><tr><td>Size of dataset</td><td>≈260 TB</td></tr><tr><td>Number of Instances</td><td>9,624,017,440</td></tr><tr><td>Number of Fields</td><td>21<sup>[1]</sup></td></tr><tr><td>Labeled Classes</td><td>N/A<sup>[2]</sup></td></tr><tr><td>Number of Labels</td><td>N/A</td></tr><tr><td>Average labels per instance</td><td>N/A</td></tr><tr><td>Algorithmic Labels</td><td>N/A</td></tr><tr><td>Human Labels</td><td>N/A</td></tr></table> <p>WebLI is built from public web pages and has 10B images without annotation. Most features are from public web pages; a few are from Cloud API<sup>[3]</sup>.</p> <p>[1] The number is for top-level features. Some of them are structured, such as “ocr_texts” etc.</p> <p>[2] The dataset is not annotated, so has no classes/labels.</p> <p>[3] <a href="#">Cloud Translation API</a> is used for text language detection and English translation; <a href="#">Cloud Vision API</a> is used for OCR detection.</p>	Size of dataset	≈260 TB	Number of Instances	9,624,017,440	Number of Fields	21 <sup>[1]</sup>	Labeled Classes	N/A <sup>[2]</sup>	Number of Labels	N/A	Average labels per instance	N/A	Algorithmic Labels	N/A	Human Labels	N/A	<p>The primary features in the dataset are <b>image</b> pixels and the associated <b>texts</b>, including alt-text, page title and OCR. Other features include:</p> <div><div>1.</div><div>Rich image and page meta information, such as URL, MIME type etc.;</div></div> <div><div>2.</div><div>Filter signals, attached to alt-text only.</div></div> <p>Additional Notes:</p> <div><div>•</div><div>No user data is included in the public data source. Personally identifiable information (PII) is filtered out during the dataset construction.</div></div>
Size of dataset	≈260 TB																	
Number of Instances	9,624,017,440																	
Number of Fields	21 <sup>[1]</sup>																	
Labeled Classes	N/A <sup>[2]</sup>																	
Number of Labels	N/A																	
Average labels per instance	N/A																	
Algorithmic Labels	N/A																	
Human Labels	N/A																	





## Sensitivity Of Data Fields

SENSITIVITY TYPE(S)	FIELD(S) WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING
User Content User Metadata User Activity Data Identifiable Data S/PII Business Data Employee Data Pseudonymous Data Anonymous Data Health Data Children’s Data <b>None</b> Others (Please Specify)	<p><b>Intentionally Collected Sensitive Data</b></p> <p>No sensitive data was intentionally collected.</p> <p><b>Unintentionally Collected Sensitive Data</b></p> <p>S/PII, pornographic images were not explicitly collected as a part of the dataset creation process.</p> <p>We used algorithmic methods and relied on other classifiers for identifying S/PII information and pornographic images, hence it is possible we may have missed some instances in the process.</p> <p>Fields that may contain such sensitive data are:</p> <ul style="list-style-type: none"><li>image pixels</li><li>alt-text, page title, OCR and their corresponding English translations</li></ul>	<p>Specifically we filtered out the following sensitive data:</p> <ol style="list-style-type: none"><li>Pornographic images using the same method as in <a href="#">ALIGN</a> and <a href="#">LIT</a>.</li><li>Text with personally identifiable information, such as email address, credit card number, phone number etc.</li></ol>

	SUPPLEMENTAL LINK(S)	RISK(S) AND MITIGATION(S)
		Like texts, images in the dataset may also contain PII (e.g. email). However, the chance for images to contain PII looks very low from our random sampling.

## Version And Maintenance

MAINTENANCE STATUS	DATASET VERSION	MAINTENANCE PLAN
<p>Regularly Updated</p> <p>New versions of the dataset have been or will continue to be made available.</p> <p>Actively Maintained</p> <p>No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.</p> <p><b>Limited Maintenance</b></p> <p><b>The data will not be updated, but any technical issues will be addressed.</b></p> <p>Deprecated</p> <p>This dataset is obsolete or is no longer being maintained.</p>	<p><b>Last Updated:</b> 2022-06-07</p> <p><b>Release Date:</b> N/A</p>	<p>Currently the dataset is in a stable version. No new features will be added unless necessary.</p> <p><b>Versioning:</b> Data is versioned following <a href="#">Semantic Versioning 2.0.0</a> (i.e., “MAJOR.MINOR.PATCH”, same as in TFDS).</p> <p><b>Updates:</b> The dataset features will be frozen and won’t allow new features unless necessary.</p> <p><b>Errors:</b> Errors will be reported and tracked.</p> <p><b>Feedback:</b> We gather feedback from researchers directly or via user groups.</p>
	NEXT PLANNED UPDATE(S)	EXPECTED CHANGE(S)
	N/A	N/A

Example Of Data Points		
PRIMARY DATA MODALITY	LINK(S) TO DATA POINT(S)	DATA FIELDS
Image Data Text Data Tabular Data Audio Data Video Data Time Series Graph Data Geospatial Data <b>Multimodal</b> Please specify Unknown Others (Please Specify)	<p>Some data point samples:</p> <div><p>(alt-text) "free stock photo of matrix and sidekick"</p><p>(OCR) "<b>card</b>", "<b>telecom</b>", "<b>5624</b>"</p><p>(page title) "Free telecom Stock Photos - Stockvault.net"</p></div> <div><p>(alt-text) "carte joyeux Noël anges et étoiles"</p><p>(OCR) "<b>joyeux Noël</b>"</p><p>(page title) "Carte Joyeux Noël !!! - anges et étoiles. - Carte postale ancienne et vue d'Hier et Aujourd'hui - Geneanet"</p></div> <div><p>(alt-text) "ทานตะวันเป็นดอกไม้ที่หันหน้าเข้าหาดวงอาทิตย์"</p><p>(page title) "ทานตะวัน ประโยชน์ดีๆ สรรพคุณเด่นๆและข้อมูลงานวิจัย"</p></div> <div><p>(alt-text) "太行山 脉 长治 太行山 大 峡谷..."</p><p>(page title) "河北太行山脉_万图壁纸网"</p></div> <p>The second image is by jopradier (<a href="#">original</a>), used under the <a href="#">CC BY-NC-SA 2.0 license</a>. Remaining images are also used with permission.</p>	<p>The TFDS format of the data fields:</p> <pre>FeaturesDict({   'alt_texts': Sequence({     'text': Text(shape=(), dtype=tf.string),     'text_en_translate': Text(shape=(), dtype=tf.string),     'text_lang': Text(shape=(), dtype=tf.string),   }),   'filters': Sequence({     'filter_id': Text(shape=(), dtype=tf.string),     'filter_signal': tf.float32,     'text_field': Text(shape=(), dtype=tf.string),     'text_index': tf.int32,   }),   'id': Text(shape=(), dtype=tf.string),   'image': Image(shape=(None, None, 3), dtype=tf.uint8),   'image_height': tf.int32,   'image_mime_type': Text(shape=(), dtype=tf.string),   'image_size': tf.int32,   'image_url': Text(shape=(), dtype=tf.string),   'image_width': tf.int32,   'ocr_texts': Sequence({     'bbox': Text(shape=(), dtype=tf.string),     'conf': tf.int32,     'lang': Text(shape=(), dtype=tf.string),     'lang_conf': tf.int32,     'rotated_bbox': Text(shape=(), dtype=tf.string),     'text': Text(shape=(), dtype=tf.string),     'text_en_translate': Text(shape=(), dtype=tf.string),     'text_lang': Text(shape=(), dtype=tf.string),   }),   'page_lang': Text(shape=(), dtype=tf.string),   'page_title': Text(shape=(), dtype=tf.string),   'page_titles': Sequence({     'text': Text(shape=(), dtype=tf.string),     'text_en_translate': Text(shape=(), dtype=tf.string),     'text_lang': Text(shape=(), dtype=tf.string),   }),   'page_url': Text(shape=(), dtype=tf.string), })</pre>

The TFDS format of the data fields:

```

FeaturesDict({
  'alt_texts': Sequence({
    'text': Text(shape=(), dtype=tf.string),
    'text_en_translate': Text(shape=(), dtype=tf.string),
    'text_lang': Text(shape=(), dtype=tf.string),
  }),
  'filters': Sequence({
    'filter_id': Text(shape=(), dtype=tf.string),
    'filter_signal': tf.float32,
    'text_field': Text(shape=(), dtype=tf.string),
    'text_index': tf.int32,
  }),
  'id': Text(shape=(), dtype=tf.string),
  'image': Image(shape=(None, None, 3), dtype=tf.uint8),
  'image_height': tf.int32,
  'image_mime_type': Text(shape=(), dtype=tf.string),
  'image_size': tf.int32,
  'image_url': Text(shape=(), dtype=tf.string),
  'image_width': tf.int32,
  'ocr_texts': Sequence({
    'bbox': Text(shape=(), dtype=tf.string),
    'conf': tf.int32,
    'lang': Text(shape=(), dtype=tf.string),
    'lang_conf': tf.int32,
    'rotated_bbox': Text(shape=(), dtype=tf.string),
    'text': Text(shape=(), dtype=tf.string),
    'text_en_translate': Text(shape=(), dtype=tf.string),
    'text_lang': Text(shape=(), dtype=tf.string),
  }),
  'page_lang': Text(shape=(), dtype=tf.string),
  'page_title': Text(shape=(), dtype=tf.string),
  'page_titles': Sequence({
    'text': Text(shape=(), dtype=tf.string),
    'text_en_translate': Text(shape=(), dtype=tf.string),
    'text_lang': Text(shape=(), dtype=tf.string),
  }),
  'page_url': Text(shape=(), dtype=tf.string),
})

```

Provenance

Data Collection & Sources

METHOD(S) USED

API  
Artificially Generated  
Crowdsourced - Paid  
Crowdsourced - Volunteer  
Vendor Collection Efforts  
**Scraped or Crawled**  
Survey, forms or polls  
Taken from other existing datasets  
Unknown  
To be determined  
**Others:**  
**Google Cloud API (see “DATASET SNAPSHOT” section)**

METHODOLOGY DETAIL(S)

**Source:** Generally accessible<sup>[1]</sup> images, alt-text and meta information from the public web  
**Is this source considered sensitive or high-risk?** [Yes / **No**]  
**Dates of Collection:** [2021-12, 2022-04]<sup>[2]</sup>  
**Primary modality of collected data:**

- Image Data
- Text Data
- Tabular Data
- Audio Data
- Video Data
- Time Series
- Graph Data
- Geospatial Data
- Unknown
- **Multimodal (Image and text)**
- Others (Please specify)

**Update Frequency for collected data:**

- Yearly / Quarterly / Monthly / Weekly / Daily / Hourly
- Static
- **Others (static then deleted to follow data policies)**

**Source:** Text language detection and English translation  
**Platform:** [Cloud Translation API](#)  
**Is this source considered sensitive or high-risk?** [Yes / **No**]  
**Dates of Collection:** [2022-05-01 to 2022-05-29]  
**Primary modality of collected data:**

- Image Data
- **Text Data**
- Tabular Data
- Audio Data
- Video Data
- Time Series
- Graph Data
- Geospatial Data
- Unknown
- Multimodal (Image and text)
- Others (Please specify)

**Update Frequency for collected data:**

- Yearly / Quarterly / Monthly / Weekly / Daily / Hourly
- Static
- **Others (static then deleted to follow data policies)**

**Source:** OCR detection  
**Platform:** [Cloud Vision API](#)  
**Is this source considered sensitive or high-risk?** [Yes / **No**]  
**Dates of Collection:** [2022-05-01 to 2022-05-29]  
**Primary modality of collected data:**

- Image Data
- **Text Data**
- Tabular Data
- Audio Data
- Video Data
- Time Series
- Graph Data
- Geospatial Data
- Unknown
- Multimodal (Image and text)
- Others (Please specify)

**Update Frequency for collected data:**

- Yearly / Quarterly / Monthly / Weekly / Daily / Hourly
- Static
- **Others (static then deleted to follow data policies)**

**Source:** robots.txt from the public web to only keep publicly available data  
**Is this source considered sensitive or high-risk?** [Yes / **No**]

SOURCE DESCRIPTION(S)

- **Web pages and images:** The publicly accessible web pages and images.
- **Cloud Translation API:** The API offered by Google Cloud to handle multilingual text with pre-trained models, functionalities including language detection and translation etc.
- **Cloud Vision API:** The API offered by Google Cloud to handle vision tasks with pre-trained models, functionalities including OCR and object detection etc.

<div><b>Dates of Collection:</b> [2022-06-03]</div> <div><b>Primary modality of collected data:</b><ul style="list-style-type: none"><li>Image Data</li><li>Text Data</li><li>Tabular Data</li><li>Audio Data</li><li>Video Data</li><li>Time Series</li><li>Graph Data</li><li>Geospatial Data</li><li>Unknown</li><li>Multimodal (Image and text)</li><li><b>Others (boolean values, i.e., whether an image/page URL is publicly available or not)</b></li></ul></div> <div><b>Update Frequency for collected data:</b><ul style="list-style-type: none"><li>Yearly / Quarterly / Monthly / Weekly / Daily / Hourly</li><li>Static</li><li><b>Others (static then deleted to follow data policies)</b></li></ul></div> <div><div>[1] Accessible to public crawlers.</div><div>[2] Most features were added in 2021-12. New features (rich meta info etc.) were added in 2022-04.</div></div>		
<b>COLLECTION CADENCE</b>	<b>DATA INTEGRATION</b>	<b>DATA PROCESSING</b>
<div>Static Data was collected once from single or multiple sources.</div> <div>Dynamic Data is updated regularly from single or multiple sources.</div> <div>Streamed Data is streamed from single or multiple sources.</div> <div>Others (static then deleted to follow data policies)</div>	<div>Web pages and images</div> <div><b>Included Fields</b> alt_texts, page_titles, page_url, page_title, page_lang; image_width, image_height, image_size, image_mime_type, image_url</div> <div><b>Excluded Fields</b> None</div> <div>Cloud Translation API</div> <div><b>Included Fields</b> text_lang, text_en_translate</div> <div><b>Excluded Fields</b> None</div> <div>Cloud Vision API</div> <div><b>Included Fields</b> ocr_texts</div> <div><b>Excluded Fields</b> None</div> <div>robots.txt</div> <div><b>Included Fields</b> None</div> <div><b>Excluded Fields</b> id (to only keep publicly available data)</div>	<div>Dataset generation</div> <div><b>Description:</b> Generates image and text pairs from the web.</div> <div><b>Methods employed:</b> Pipelines to read web pages and images from the public web and store needed info as structured data.</div> <div><b>Tools or libraries:</b> Pipelines</div> <div>Text language detection and translation</div> <div><b>Description:</b> Detects text language and translates to English.</div> <div><b>Methods employed:</b> A pipeline to call Cloud Translation API, and add language detection and translation results to the dataset.</div> <div><b>Tools or libraries:</b> <a href="#">Cloud Translation API</a></div> <div>Publicly available data filtering</div> <div><b>Description:</b> Only keep publicly available data.</div> <div><b>Methods employed:</b> Access public robots.txt to determine which web pages and images are visible to all crawlers, and keep those publicly available web pages.</div> <div><b>Tools or libraries:</b> robots.txt</div>
<b>Collection Criteria</b>		
<b>DATA SELECTION</b>	<b>DATA INCLUSION</b>	<b>DATA EXCLUSION</b>
<b>Public web pages:</b> The dataset is built from public web pages.	Records that are not excluded are in the final dataset.	<div><b>No adult content</b><sup>[1]</sup>: Images which are identified as having adult content are excluded.</div> <div><b>No sensitive text</b><sup>[1]</sup>: Texts (alt-text, page title and OCR) which are identified as PII are excluded.</div> <div><b>No empty text</b><sup>[2]</sup>: Empty texts (alt-text, page title and OCR) are excluded.</div> <div><b>Image shape and paired-text frequency filtering.</b></div> <div><div>[1] See “<a href="#">Sensitivity Of Data Fields</a>” for filtering details.</div><div>[2] Text which only has \n, \t, \r, and blank characters.</div></div>



Human Attributes			
<b>SENSITIVE HUMAN ATTRIBUTE(S)</b>	<b>INTENTIONALITY</b>	<b>RATIONALE</b>	



Race Gender Ethnicity Socio-economic status Geography Language Sexual Orientation Religion Age Culture Disability Experience or Seniority Others (Please Specify) <b>None</b>	<b>Intentionally Collected Attributes</b>  No human attributes were labeled or collected as a part of the dataset creation process.  <b>Unintentionally Collected Attributes</b>  Human attributes were not explicitly collected as a part of the dataset creation process but may be inferred using additional methods from the image pixels or the text fields. For instance, a gender classifier can be used to annotate the data with gender labels.	N/A (no human attributes were labeled or collected)
--	--	---

Motivations & Intentions		
Intended Use		
DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
Safe for production use <b>Safe for research use</b> Conditional use - some unsafe applications Only approved use Others (Please Specify)	<ul style="list-style-type: none"> <li><b>Pre-train vision-language backbones:</b> including both discriminative and generative models.</li> <li><b>Pre-train vision-only backbones:</b> e.g. using contrastive learning to train vision models.</li> </ul>	The dataset might not be suitable for pre-training language-only backbones (e.g. BERT) due to its noisy text.
	RESEARCH OR PROBLEM SPACE(S)	CITATION GUIDELINE(S)
	<p>The dataset is a large-scale multilingual vision-language dataset, and suitable for pre-training strong backbones for both single-modality and cross-modality tasks.</p> <p>Vision-only tasks:</p> <ul style="list-style-type: none"> <li>image classification (e.g. ImageNet)</li> <li>object detection (e.g. COCO, LVIS)</li> <li>segmentation of instance/panoptic/semantics/... (e.g. COCO)</li> </ul> <p>Language-only tasks:</p> <ul style="list-style-type: none"> <li>question answering (e.g. XQUAD)</li> <li>natural language inference (e.g. XNLI)</li> </ul> <p>Vision+Language tasks:</p> <ul style="list-style-type: none"> <li>image captioning (e.g. COCO Captions, CC3M)</li> <li>image-text retrieval (e.g. COCO Captions)</li> <li>visual question answering (e.g. VQAv2)</li> <li>zero-shot transfer vision tasks (e.g. ImageNet)</li> </ul> <p>Multilingual tasks:</p> <ul style="list-style-type: none"> <li>image captioning (e.g. Crossmodal-3600)</li> <li>image-text retrieval (e.g. Crossmodal-3600)</li> </ul>	Please cite this paper.