W9 Labeling Instructions
In order to do fair and thorough evaluation of the model, the evaluation set should take the following dimensions into consideration

1. **Revisions**: Documents should be in English for both (Rev. 10-2018, Rev. 11-2017) revisions.
2. **Size**: It is recommended to have 500+ documents for each evaluation set.
3. **Template Diversity**: The documents should be diverse, e.g. different templates, issue year, different sources, states (if applicable) ... etc.
4. **Value Diversity**: Documents should capture all possible values. E.g. documents should cover all checkbox options for “Federal Tax Classification”. Same applies to all ENUM or optional entities.
5. **Handwritten**: It is preferred to separate handwritten documents from typed ones (i.e. have 2 separate evaluation sets) and report metrics for each.
Processor Version

These instructions are specific for W9 Processor version:

“Google Release Candidate 2022-06-23”
# Entity List

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Base Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>business_name</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>address</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>city_state_zip</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>account_numbers</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>ltc_tax_classification</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>federal_tax_classification_other</td>
<td>Text field that is usually populated when &quot;Other&quot; checkbox option is selected.</td>
<td>string</td>
</tr>
<tr>
<td>exempt_payee_code</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>exempt_fatca_code</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>ssn</td>
<td>Value should be composed of digits only.</td>
<td>number</td>
</tr>
<tr>
<td>ein</td>
<td>Value should be composed of digits only.</td>
<td>number</td>
</tr>
<tr>
<td>has_signature</td>
<td>Enum: YES, NO (Case Sensitive)</td>
<td>string (enum)</td>
</tr>
<tr>
<td>has_signature_date</td>
<td>Enum: YES, NO (Case Sensitive)</td>
<td>string (enum)</td>
</tr>
<tr>
<td>form_revision_date</td>
<td></td>
<td>string</td>
</tr>
<tr>
<td>federal_tax_classification</td>
<td>Checkbox with 7 options one of them should be selected.</td>
<td>string (enum)</td>
</tr>
</tbody>
</table>

*Note: There will be zero or one instance of each entity.*
General Labeling Instructions
Occurrence Type

There are 4 occurrence types which restricts the number of annotations for each entity:

- **OPTIONAL_ONCE**: Entity can be present zero or multiple times but should have one unique value.
- **OPTIONAL_MULTIPLE**: Entity will appear zero or multiple times and with different values.
- **REQUIRED_ONCE**: Entity can be present once or multiple times but should have one unique value.
- **REQUIRED_MULTIPLE**: Entity will appear once or multiple times and with different values.

All W9 form entities belongs to any of the occurrence type mentioned above.
Optical Character Recognition (OCR) Errors

Drawing bounding boxes around entities will automatically populate a text captured through OCR (Optical Character Recognition). In some cases, the captured text is not accurate and doesn’t reflect what is on the page, especially for documents that are blurry or of low resolution. These values should be corrected manually through “Value” input text.
Bounding Box Annotation

Bounding boxes should fit the correct text as much possible:

- Avoid overlapping bounding boxes. This can be tricky when values are close to each other.
- Avoid oversized boxes which incorporates background artifacts such as table borders.

Example 1
Incorrect Annotation: Bounding boxes are overlapping

Example 2
Correct Annotation

Example 3
Incorrect Annotation: Bounding boxes oversized
Case Sensitivity

DocAI evaluation is case sensitive and can affect the final evaluation. While annotation please be aware to fix any casing issues generated by OCR.

Please note that ENUM values are also case sensitive and the values should be fixes as listed in the Entity List slide.

Examples:

- “John doe” != “John Doe”
- “Ccorporation” != “CCorporation”
Numbers

Numbers and Account Numbers such as SSN and EIN should only be composed of digits. In the example below, the OCR captured extra dashes and spaces that should be manually removed.
Money

For money entries the annotation bounding box and value text should include the currency if present on the document.

If the document contains currency with no values, then annotations could be omitted, as shown in the example on the bottom right.
Empty Values

In general, if the value does not exist in the form then there is no need to label it. In the example below, `ltc_tax_classification` should not be annotated since there is no value present in the designated field.

It is also possible to annotate empty entities (with empty values) but that will not affect the final metrics.

Note: There is one exception to this rule, and will be explained in the next slide.
YES/NO Values

When annotating ENUM types that includes YES/NO options such as `has_signature` and `has_signature_date`, a bounding box must be drawn at that field and the value should be changed to either “YES” if the value exists or “NO” if it does not (Note: these values are case sensitive). This will allow the groundtruth values to match the processor predicted output. Examples:

Example 1: YES

Example 2: NO
Checkbox entities are tricky to annotate. The annotation required for evaluation is different than the annotation for training.

For Evaluation: Checkbox annotation needs to be drawn around the option that is selected. Then the Value should be changed to one of the ENUM options listed in the Field Entity List slide (the value is case sensitive).

Example 1: Other option is selected

Example 2: Trust/estate option is selected
For training, checkbox entities should be added to the dataset schema first by following the steps:

1. Click on “Edit Schema” button.
2. Click on “Create Label”
3. Type a suitable name, change “Data Type” to “Checkbox” and click “Create”.
4. Repeat for all checkbox options.
During annotation for training, all checkbox options should be labeled with the correct value as shown in the example below: