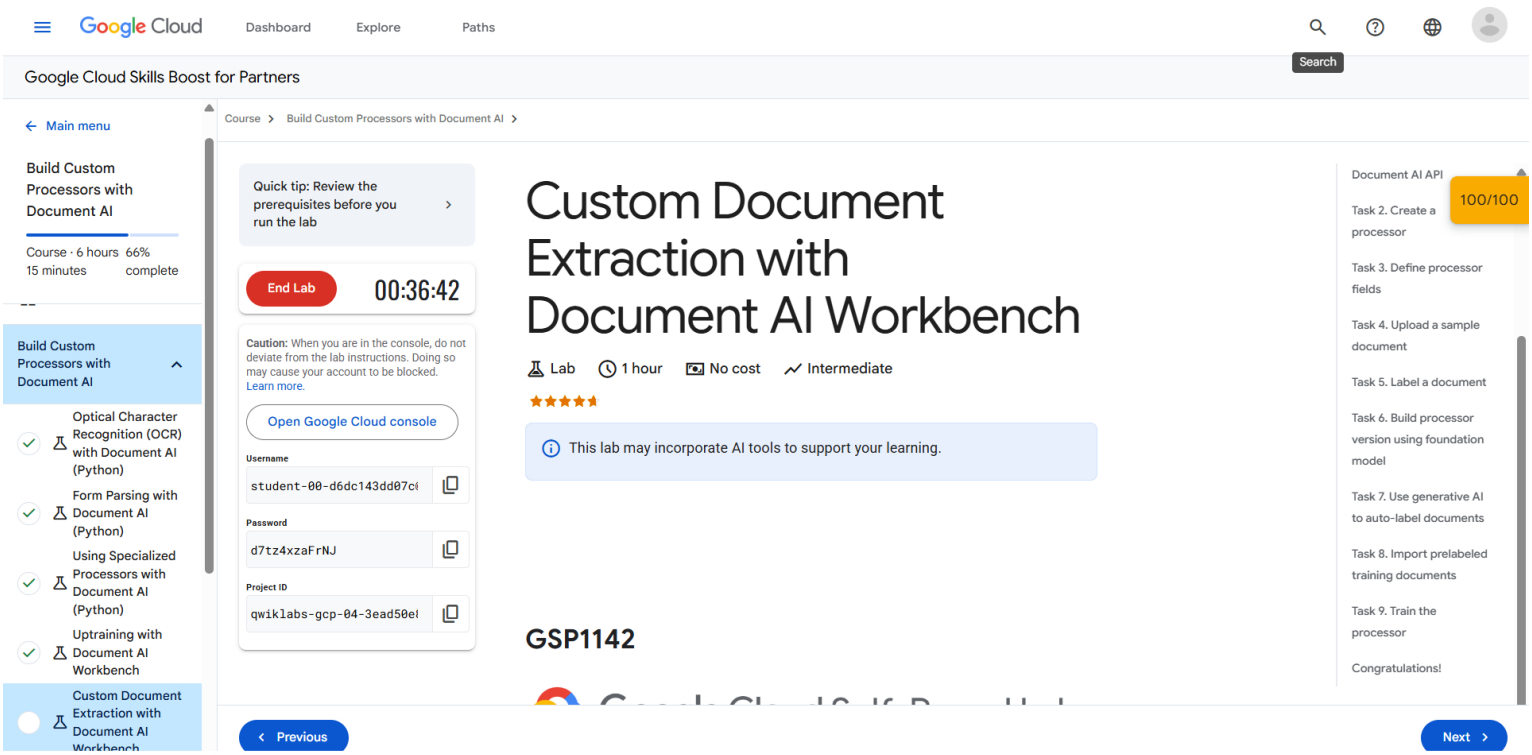


Overview

Document AI is a document understanding solution that takes unstructured data (e.g. documents, emails, invoices, forms, etc.) and makes the data easier to understand, analyze, and consume. The API provides structure through content classification, entity extraction, advanced searching, and more. With Document AI Workbench, you can achieve higher document processing accuracy by creating fully customized models using your own training data.

You can create Custom Document Extractors (CDE) that are specifically suited to your documents, and trained and evaluated with your data. This processor identifies and extracts entities from your documents. You can then use this trained processor on additional documents. You typically would use a CDE on documents that are all of one type, such as your institution's enrollment forms.

Custom Document Extractor that processes W-2 (US tax form) documents. Most of the document preparation work has been done so that you can focus on the other mechanics of creating a CDE.

Objectives

In this lab, you will learn how to perform the following tasks:

- Create a Custom Document Extractor in Document AI Workbench
- Define and create the processor schema
- Import documents
- Annotate documents manually in Document AI Workbench
- Use generative AI to auto-label documents
- Kick off a training job for the processor

Setup and requirements

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which

starts when you click **Start Lab**, shows how long Google Cloud resources are made available to you.

This hands-on lab lets you do the lab activities in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials you use to sign in and access Google Cloud for the duration of the lab.

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).

This prevents conflicts between your personal account and the student account, which may cause extra charges incurred to your personal account.

- Time to complete the lab—remember, once you start, you cannot pause a lab.

Note: Use only the student account for this lab. If you use a different Google Cloud account, you may incur charges to that account.

How to start your lab and sign in to the Google Cloud console

1. Click the **Start Lab** button. If you need to pay for the lab, a dialog opens for you to select your payment method. On the left is the Lab Details pane with the following:
    - The Open Google Cloud console button
    - Time remaining
    - Other information, if needed, to step through this lab
2. Click **Open Google Cloud console** (or right-click and select **Open Link in Incognito Window** if you are running the Chrome browser).

    The lab spins up resources, and then opens another tab that shows the Sign in page.

    Tip: Arrange the tabs in separate windows, side-by-side.

    Note: If you see the **Choose an account** dialog, click **Use Another Account**.
3. If necessary, copy the **Username** below and paste it into the **Sign in** dialog.

    ```
    student-00-d6dc143dd07c@qwiklabs.net
    ```

    You can also find the Username in the Lab Details pane.
5. Copy the **Password** below and paste it into the **Welcome** dialog.

    ```
    d7tz4xzaFrNJ
    ```

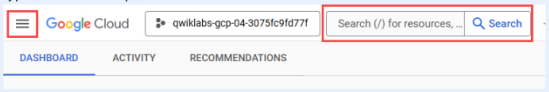
    You can also find the Password in the Lab Details pane.
6. Click **Next**.

    Important: You must use the credentials the lab provides you. Do not use your Google Cloud account credentials.

    Note: Using your own Google Cloud account for this lab may incur extra charges.
7. Click through the subsequent pages:
    - Accept the terms and conditions.
    - a temporary account).
    - Do not sign up for free trials.

After a few moments, the Google Cloud console opens in this tab.

Note: To access Google Cloud products and services, click the **Navigation menu** or type the service or product name in the **Search** field.



Activate Cloud Shell

Cloud Shell is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud. Cloud Shell provides command-line access to your Google Cloud resources.

2. Click through the following windows:
    - Continue through the Cloud Shell information window.
    - Authorize Cloud Shell to use your credentials to make Google Cloud API calls.

When you are connected, you are already authenticated, and the project is set to your **Project_ID**, `qwiklabs-gcp-04-3ead50e8fb32`. The output contains a line that declares the **Project_ID** for this session:

```
Your Cloud Platform project in this session is set to qwiklabs-gcp-04-3ead
```

`gcloud` is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.

3. (Optional) You can list the active account name with this command:

4. Click **Authorize**.

Output:

```
ACTIVE: *
ACCOUNT: student-00-d6dc143dd07c@qwiklabs.net

To set the active account, run:
    $ gcloud config set account `ACCOUNT`
```

5. (Optional) You can list the project ID with this command:

```
gcloud config list project
```

Output:

```
[core]
```

Note: For full documentation of `gcloud`, in Google Cloud, refer to the gcloud CLI overview guide.

Task 1. Enable the Document AI API

Before you can begin using Document AI, you must enable the API.

1. In Cloud Shell, run the following command to enable the API for Document AI.

```
gcloud services enable documentai.googleapis.com
```

You should see something like this:

```
Operation "operations/..." finished successfully.
```

2. Run the following command to install the Python client libraries for Document AI.

```
pip3 install --upgrade google-cloud-documentai
```

You should see something like this:

```
...
Installing collected packages: google-cloud-documentai
Successfully installed google-cloud-documentai-2.15.0
```

Now, you're ready to use the Document AI API!

Enable the Document AI API

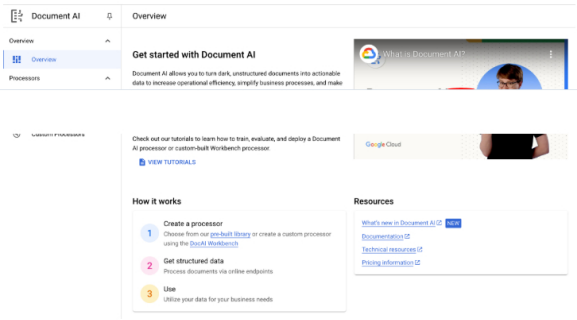
Check my progress

Task 2. Create a processor

You must first create an Custom Document Extractor processor to use for this lab.

You must first create a Form Parser processor instance to use in the Document AI Platform for this tutorial.

1. From the **Navigation menu** select **View all products**. Under **Artificial Intelligence**, select **Document AI**.



2. Click **Create Custom Processor**.
3. Inside the **Custom Extractor** box, click **Create Processor**.
4. Give it the name `lab-custom-extractor` and select the region **US (United States)** on the list.
5. Click **Create** to create your processor.

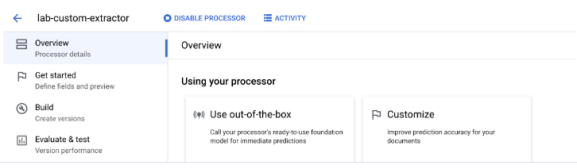
Click **Check my progress** to verify the objective.

Check my progress

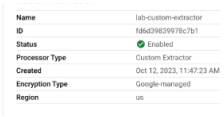
Assessment Completed!

Task 3. Define processor fields

You are now on the **Processor overview** page of the processor you just created.



Basic information

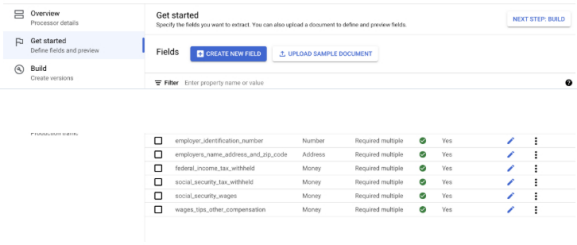


You can specify the fields you want the processor to extract and begin labeling documents.

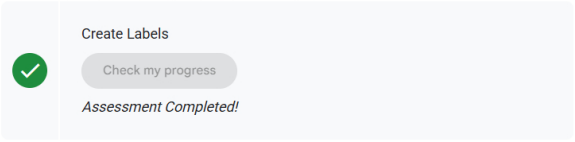
1. Click on the **Get started** tab. The **Fields** menu appears.
2. Click **Create New Field**.
3. Enter the name for the field. Select the **Data type** and the **Occurrence**. Click **Create**. Refer to Define processor schema for detailed instructions on creating and editing a schema.
4. Create each of the following labels for the processor schema.

Name	Data Type	Occurrence
`employees_social_security_number`	Number	Required multiple
`employer_identification_number`	Number	Required multiple
`employers_name_address_and_zip_code`	Address	Required multiple
`federal_income_tax_withheld`	Money	Required multiple
`social_security_tax_withheld`	Money	Required multiple
`social_security_wages`	Money	Required multiple
`wages_tips_other_compensation`	Money	Required multiple

You can also create and use other types of labels in your processor schema, such as checkboxes and tabular entities. For example, the W-2 forms contain a **Statutory employee**, **Retirement plan**, and **Third party sick pay** check boxes that you could also add to the schema.



Click **Check my progress** to verify the objective.



Task 4. Upload a sample document

1. Click **Upload Sample Document**.
2. In the sidebar, click **Import documents from Google Cloud Storage**.
3. For this example, enter this bucket name in Source path. This links directly to one document.

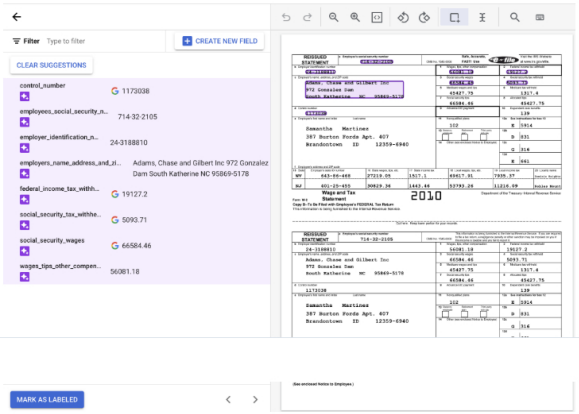
```
cloud-samples-
data/documentai/Custom/W2/PDF/W2_XL_input_clean_2950.pdf
```

4. Click **Import**.

You are redirected to the labeling console.

The process of selecting text in a document, and applying labels is known as annotation.

1. When you're at the labeling console, notice that many of the labels are already populated.



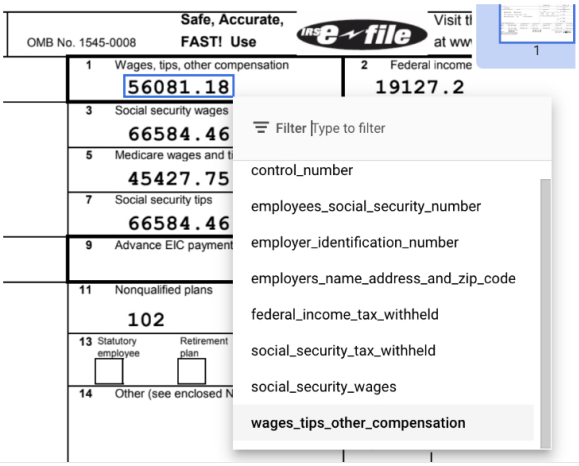
Note: Your results might look slightly different than the sample image.

2. To use the suggested labels, hold the pointer over each label in the side panel, and click on the check mark to confirm the label is correct. You can edit the values if they do not match the document text.

3. In this example, the values at the bottom of the document were not identified automatically, so you will need to label them manually.

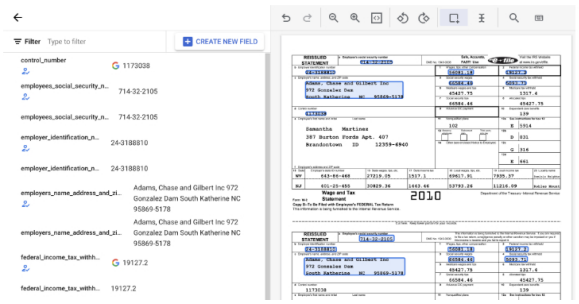
4. Use the **Bounding box** tool by default, or the **Select text** tool for multi-line values, to select the content and apply the label.

Note: The **Select text** tool does not work for all text values, so use the **Bounding box** if appropriate. You can also select non-text fields such as checkboxes using the **Bounding box** tool.



6. Review the detected text values to ensure that they reflect the correct text from the document.

The labeled W-2 document should look like this when complete:



7. If needed, you can click **Create New Field** to add a new field to the schema from this page.

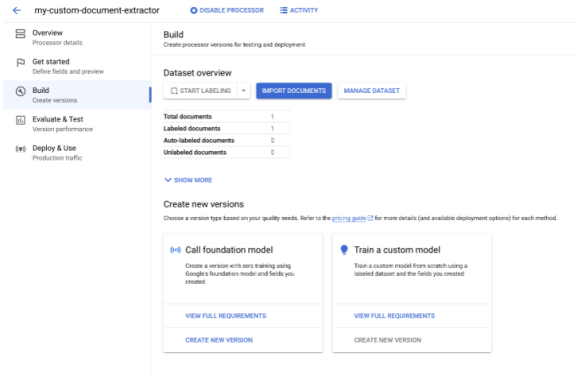
8. Click **Mark as Labeled** when you have finished annotating the document.

You are redirected to the **Get started** tab.

Task 6. Build processor version using foundation model

After labeling a single document, you can create a processor version using the pretrained foundation model to extract entities.

1. Click on the **Build** tab.



2. Under **Call foundation model**, click **Create New Version**.

3. Enter a name for your processor version, such as `w2-foundation-model`.

Note: Once you create a processor version, you cannot delete fields you have created. You can disable them on the fields page if you no longer need them.

5. Optional: Click on the **Deploy & Use** tab. On this page, you can view the available processor versions and the deployment status of the new version.

You test and evaluate this version later in the lab.

Click **Check my progress** to verify the objective.

Build processor version using foundation model

Check my progress

Assessment Completed!

Task 7. Use generative AI to auto-label documents

The foundation model can accurately extract fields for a variety of document types, but you can also provide additional training data to improve the accuracy of the model for specific document structures.

Document AI Workbench uses the label names you define and previous annotations to make it quicker and easier to label documents at scale with auto-labeling.

1. Go to the **Build** page.

2. Click **Import Documents**.

3. In the sidebar, click **Import documents from Google Cloud Storage**.
4. Enter this bucket name in **Source path**. This contains unlabeled W-2 PDF files.

```
[illegible]
```

5. From the **Data split** list, select **Auto-split**. This automatically splits the documents to have 80% in the training set and 20% in the test set.
6. In the **Auto-labeling** section, select the **Import with auto-labeling** checkbox.
7. Select the foundation model processor version you just created to label the documents.
8. Click **Import** and wait for the documents to import. You can leave this page and return later.
9. You must verify the auto-labeled documents before you can use them for training or testing. Click **Start Labeling** to view the auto-labeled documents.
10. To use the suggested labels, hold the pointer over each annotation, and click on the check mark to confirm the label is correct. You can edit the values if they do not match the document text.
11. Click **Mark as Labeled** when you have finished annotating the document.
12. Repeat for each auto-labeled document. For this tutorial, you can skip any

Task 8. Import prelabeled training documents

In this lab, you are provided with prelabeled data. If working on your own project, you have to determine how to label your data. Refer to Labeling options for more details. In general, more training data produces higher accuracy.

1. Go to the **Build** page.
2. Click **Import Documents**.
3. In the sidebar, click **Import documents from Google Cloud Storage**.
4. Enter the following path in **Source path**. This bucket contains prelabeled documents in the Document .JSON format.

```
cloud-samples-data/documentai/Custom/W2/JSON-2
```

5. From the **Data split** list, select **Auto-split**. This automatically splits the documents to have 80% in the training set, and 20% in the test set. Leave **Import with auto-labeling** unchecked.
6. Click **Import**. Import takes several minutes.
7. (Optional) From the **Build** page, you can access the **Manage Dataset** console to view and edit all documents and labels in the dataset.

Task 9. Train the processor

Now that you have sufficient training and test data, you can train the processor.

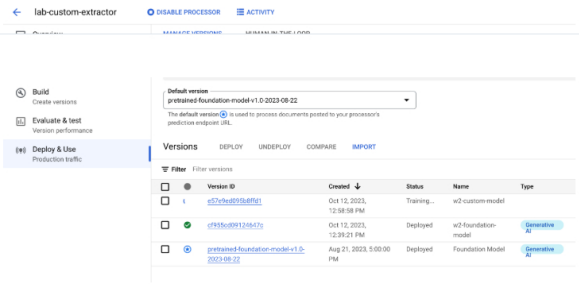
1. Under **Train a custom model**, click **Create New Version**.

   If **Create New Version** cannot be clicked, click on **View Full Requirements** for information about the dataset requirements.
2. In the **Version name** field, enter a name for this processor version, such as `w2-custom-model`.
3. (Optional) Click **View Label Stats** to find information about the document labels. That can help determine your coverage. Click **Close** to return to the training setup.

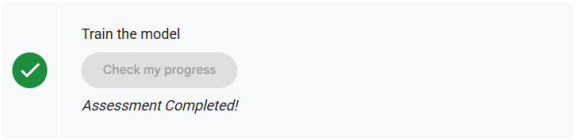
4. Under **Model training method**, select **Model based**.

5. Click **Start training**.

6. (Optional) Click on the **Deploy & Use** tab. On this page, you can view the available processor versions and the training status of the new version.



Click **Check my progress** to verify the objective.



Great! You have now started training your first Custom Document AI Processor. Since the training job will take around a few hours, the lab will end here. If you are interested in

Congratulations!

Congratulations, in this lab you've successfully used Document AI to create a Custom Document Extraction processor, import a dataset, and label example documents. You can now use this processor to parse documents in this format just as you would for any Specialized Processor. You can also use this processor to label new documents using auto-labeling as well as use the Document AI Workbench to manage your training data and training jobs.

Next steps / Learn more

Check out the following resources to learn more about Document AI and the Python

- The Future of Documents - YouTube Playlist
- Document AI Documentation
- Document AI Python Client Library
- Document AI Samples

Google Cloud training and certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.

Manual Last Updated March 17, 2025

Lab Last Tested March 17, 2025