

Create bucket and upload files

The screenshot shows the 'Create a bucket' wizard in the Google Cloud console. The left sidebar contains 'Cloud Storage' with sub-items 'Buckets', 'Monitoring', and 'Settings'. The main content area is titled 'Create a bucket' and includes a back arrow and a forward arrow. The wizard consists of several steps: 1. 'Name your bucket' with a text input field containing 'Ex. 'example', 'example_bucket-1', or 'example.com'' and a 'CONTINUE' button. 2. 'Choose where to store your data' showing 'Location: us (multiple regions in United States)' and 'Location type: Multi-region'. 3. 'Choose a storage class for your data' showing 'Default storage class: Standard'. 4. 'Choose how to control access to objects' showing 'Public access prevention: On' and 'Access control: Uniform'. 5. 'Choose how to protect object data' showing 'Soft delete policy: Enabled', 'Object versioning: Disabled', 'Bucket retention policy: Disabled', 'Object retention: Disabled', and 'Encryption type: Google-managed'. At the bottom are 'CREATE' and 'CANCEL' buttons. On the right, a 'Good to know' section includes 'Location pricing' with a table showing storage costs and an 'ESTIMATE YOUR MONTHLY COST' link.

Google Cloud herbaria-ai document ai workbench Search

Cloud Storage Create a bucket

- Name your bucket**
Pick a globally unique, permanent name. [Naming guidelines](#)
Ex. 'example', 'example_bucket-1', or 'example.com'
Tip: Don't include any sensitive information
LABELS (OPTIONAL)
[CONTINUE](#)
- Choose where to store your data**
Location: us (multiple regions in United States)
Location type: Multi-region
- Choose a storage class for your data**
Default storage class: Standard
- Choose how to control access to objects**
Public access prevention: On
Access control: Uniform
- Choose how to protect object data**
Soft delete policy: Enabled
Object versioning: Disabled
Bucket retention policy: Disabled
Object retention: Disabled
Encryption type: Google-managed
[CREATE](#) [CANCEL](#)

Good to know

Location pricing
Storage rates vary depending on the storage class of your data and location of your bucket. [Pricing details](#)

Current configuration: Multi-region / Standard

Item	Cost
us (multiple regions in United States)	\$0.026 per GB-month
With default replication	\$0.020 per GB written

[ESTIMATE YOUR MONTHLY COST](#)

The screenshot shows the 'Bucket details' page for a bucket named 'herbaria-test-images2'. The left sidebar is the same as the previous screenshot. The main content area is titled 'Bucket details' and includes a back arrow, a 'REFRESH' button, and a 'LEARN' button. The bucket name 'herbaria-test-images2' is displayed. Below the name, a table shows bucket properties: Location (us-east1 (South Carolina)), Storage class (Standard), Public access (Not public), and Protection (None). A tabbed interface shows 'OBJECTS' as the active tab, with other tabs including 'CONFIGURATION', 'PERMISSIONS', 'PROTECTION', 'LIFECYCLE', 'OBSERVABILITY', and 'INVENTORY REPORTS'. Below the tabs, there's a breadcrumb 'Buckets > herbaria-test-images2' and a set of action buttons: 'UPLOAD FILES', 'UPLOAD FOLDER' (highlighted with a red box), 'CREATE FOLDER', 'TRANSFER DATA', 'MANAGE HOLDS', 'EDIT RETENTION', 'DOWNLOAD', and 'DELETE'. Below these buttons is a filter section with 'Filter by name prefix only' and a 'Filter' button. A table lists the objects in the bucket, showing one folder named 'temp_images/'. The table has columns for Name, Size, Type, Created, Storage class, Last modified, Public access, and Version history.

Google Cloud herbaria-ai document ai workbench Search REFRESH LEARN

Cloud Storage Bucket details

herbaria-test-images2

Location	Storage class	Public access	Protection
us-east1 (South Carolina)	Standard	Not public	None

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS

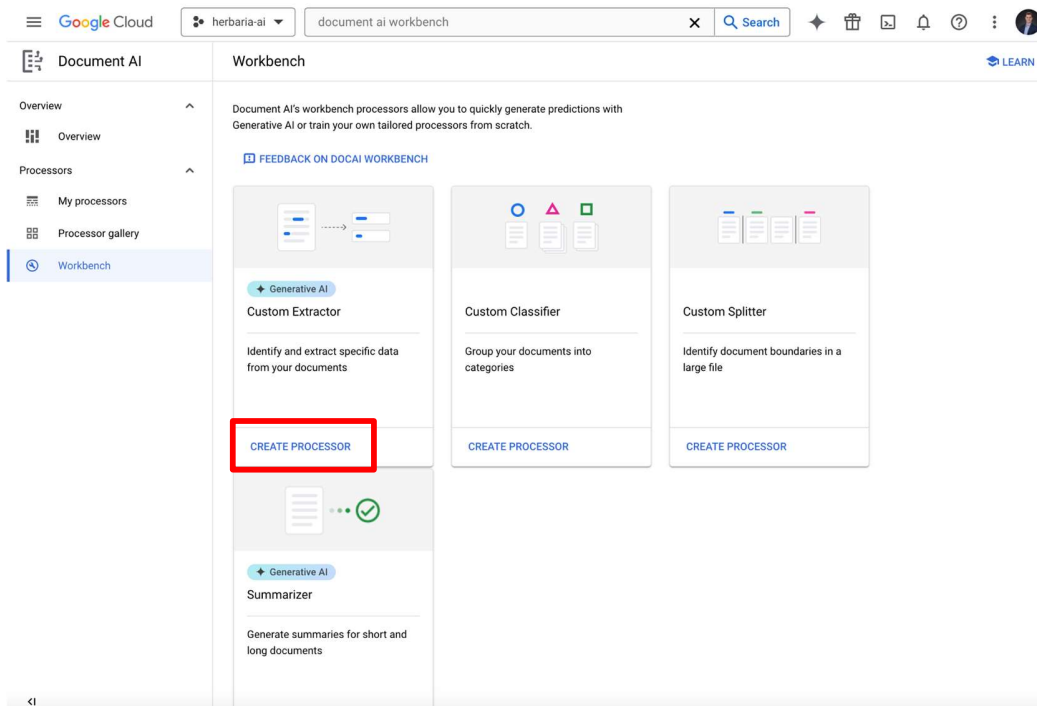
Buckets > herbaria-test-images2

[UPLOAD FILES](#) [UPLOAD FOLDER](#) [CREATE FOLDER](#) [TRANSFER DATA](#) [MANAGE HOLDS](#) [EDIT RETENTION](#) [DOWNLOAD](#) [DELETE](#)

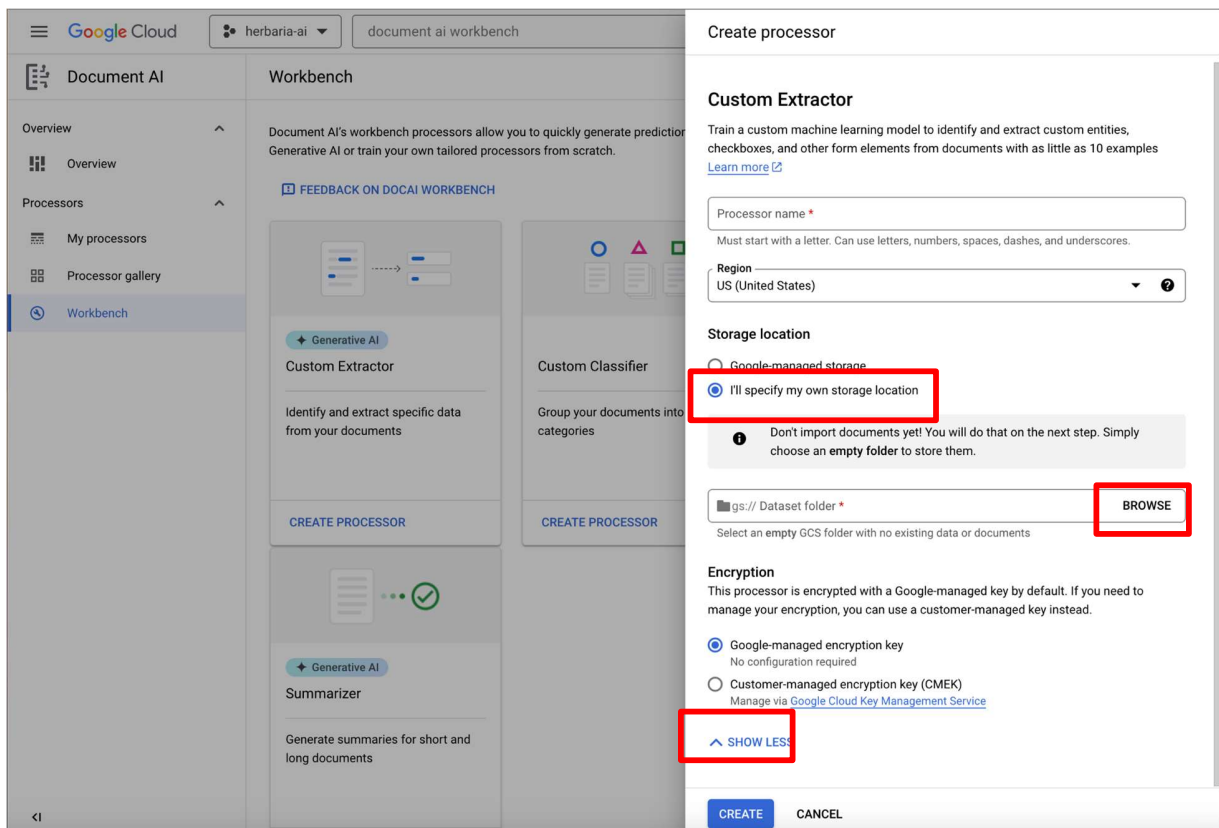
Filter by name prefix only Filter objects and folders Show Live objects only

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history
temp_images/	—	Folder	—	—	—	—	—

Open Document AI Workbench and Create a Processor

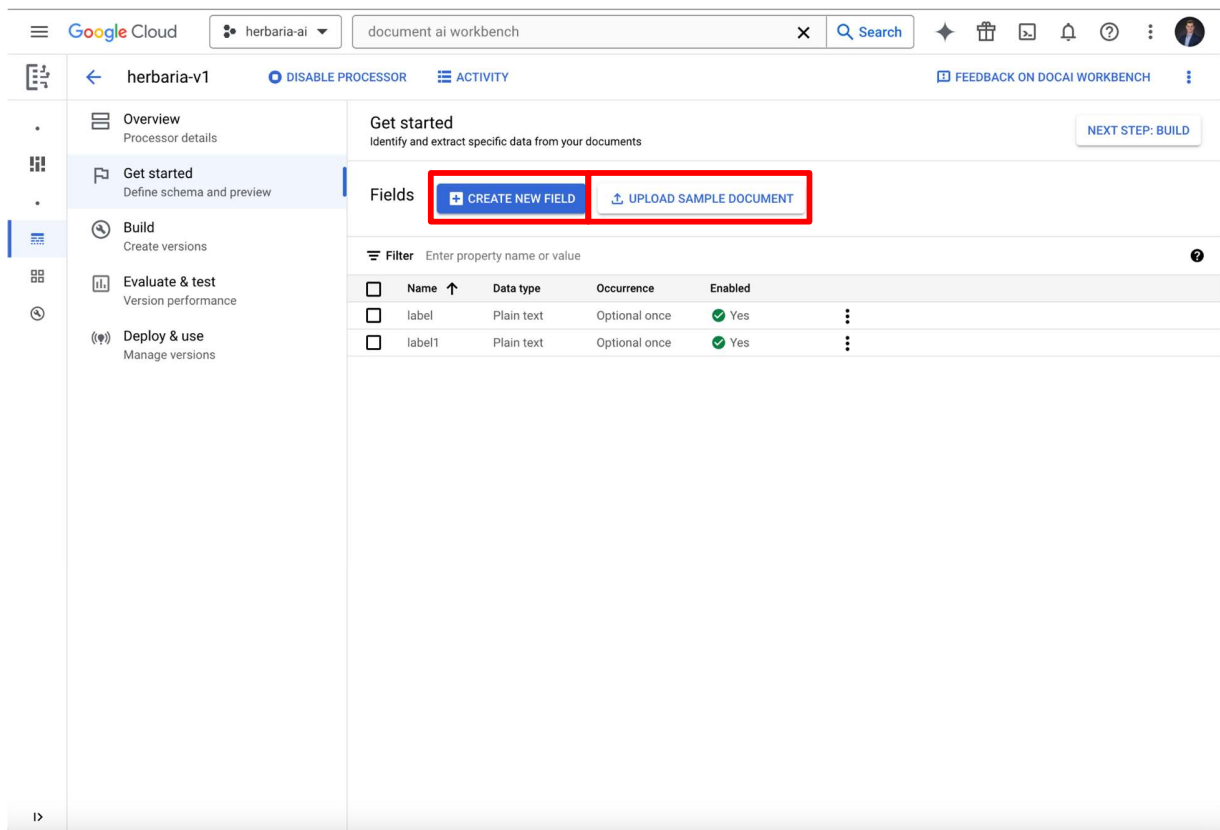


Specify the Storage Location you would like to use for results (must be an empty folder)



Create Labels and Upload Training/Test Documents

Labels can be named and you can choose the number of possible occurrences they can have. Google uses Generative AI to place bounding boxes on the document (see example below) based on a sample document you can upload. We will further explore how detailed we can be with these labels and how flexible they are to different document layouts



Google Cloud herbaria-ai document ai workbench

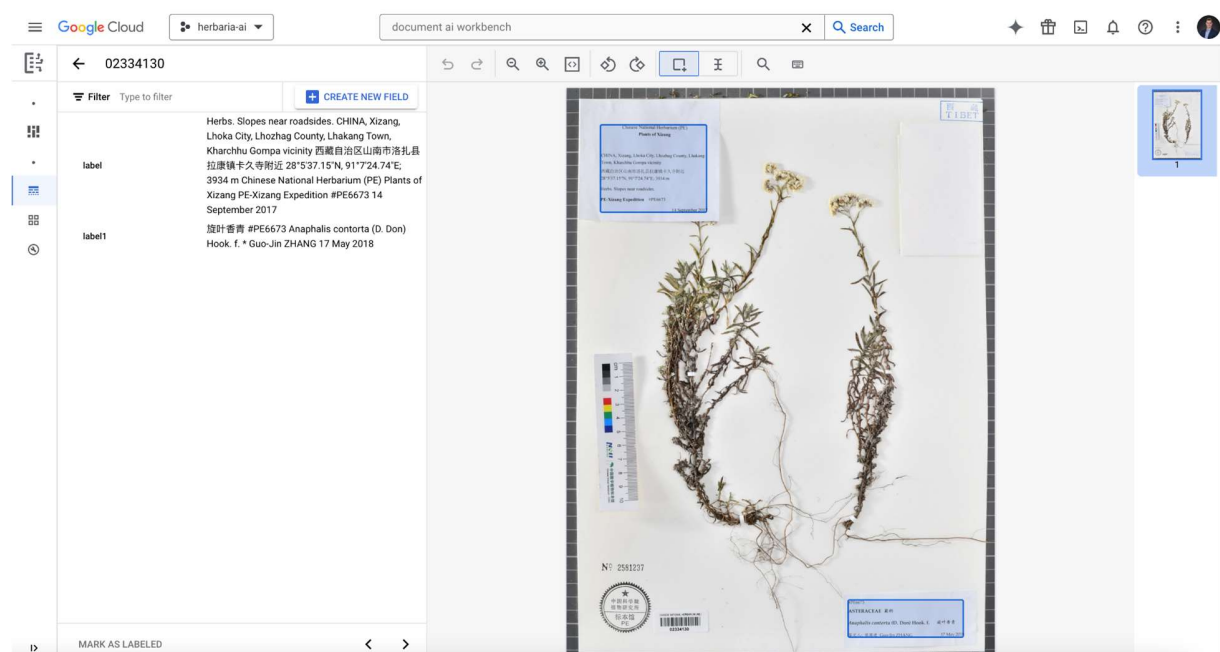
herbaria-v1 [DISABLE PROCESSOR](#) [ACTIVITY](#) [FEEDBACK ON DOCAI WORKBENCH](#)

Get started
Identify and extract specific data from your documents [NEXT STEP: BUILD](#)

Fields [CREATE NEW FIELD](#) [UPLOAD SAMPLE DOCUMENT](#)

Filter Enter property name or value

<input type="checkbox"/>	Name	Data type	Occurrence	Enabled	
<input type="checkbox"/>	label	Plain text	Optional once	Yes	⋮
<input type="checkbox"/>	label1	Plain text	Optional once	Yes	⋮



Google Cloud herbaria-ai document ai workbench

02334130 [CREATE NEW FIELD](#)

Filter Type to filter

label
Herbs. Slopes near roadsides. CHINA, Xizang, Lhoka City, Lhokha County, Lhakang Town, Kharchhu Gumpa vicinity 西藏自治区山南市洛扎县拉康镇卡久寺附近 28°53'15"N, 91°7'24.74"E, 3934 m Chinese National Herbarium (PE) Plants of Xizang PE-Xizang Expedition #PE6673 14 September 2017

label1
旋叶香青 #PE6673 Anaphalis contorta (D. Don) Hook. f. * Guo-Jin ZHANG 17 May 2018

MARK AS LABELED

Import train/test documents and label

You can then import documents to train the model and test your labeling. You can either specify a standard train test split or upload all documents as training data and later move a specific number of them to test data. You can either auto-label the documents you upload based on the sample labeled document you created or you can label by hand certain documents.

The screenshot shows the 'Manage Dataset' page in the Google Cloud AI Workbench. The left sidebar contains a list of dataset categories: All documents (178), Labeled (5), Auto-labeled (173), Unlabeled (0), Suggested (2) with a 'NEW' badge, and an 'IMPORT DOCUMENTS' button highlighted with a red box. Below this is a 'Data split' section with Training (158), Test (20), and Unassigned (0). A 'Filter' bar is present above the document thumbnails. The main area displays a grid of document thumbnails, each with a label ID (e.g., 02334038, 02334010, 02334004, 02333964, 02333970, 02333971) and a '1 page' indicator. On the right, there are three sections: 'Training' with a 'TRAIN NEW VERSION' button, 'Auto-label documents' with an 'AUTO-LABEL DOCUMENTS' button highlighted by a red box, and 'Label stats' with a 'VIEW LABEL STATS' button. At the bottom right, there is an 'Export dataset' section with an 'EXPORT DATASET' button.

Deployment

Once your model is trained and evaluated it can then be deployed using the processor's prediction endpoint URL. The model can of course be retrained and reevaluated at any time as new documents are added or as labels are added or modified.

The screenshot shows the 'Manage Versions' page in the Google Cloud AI Workbench for the 'herbaria-v1' processor. The left sidebar shows the navigation menu with 'Deploy & use' selected. The main area displays the 'Default version' as 'pretrained-foundation-model-v1.0-2023-08-22'. Below this, there is a table of versions. The table has columns for Version ID, Created, Status, Name, Type, F1 score, and API. The first row shows a version created on Aug 21, 2023, at 8:00:00 PM, with a status of 'Deployed', name 'Google Stable', type 'Generative AI', and an F1 score of 0.0. The table also includes links for 'VIEW DETAILS' and 'SAMPLE REQUEST'. At the top right, there is a 'FEEDBACK ON DOCAI WORKBENCH' link.

Next Steps

We will continue finding ways to streamline this pipeline and improve the results of the labeling and training process. We have also enabled the Google Translation API and will look into implementations of it within Document AI itself. As well as test calling our processor in a Python notebook.

