**Tagging Tool Instruction Manual**

Project: JLL DocAI POC

Date: 08/16/2021

Authors: Quantiphi

Prepared for: Taggers

Document type: Tagging Tool Instruction Manual

Document Version: 1

Contents

[**1. Introduction**](#_ya5hp73cqxhr) **3**

[**2. Procedure**](#_j7aapijk4wju) **3**

[2.1 Step 1 - Access to the GCP console](#_x4vyop5x65g) 3

[2.1.1 Select a project](#_gnkclzwsmge8) 3

[2.1.2 Select Natural Language Product](#_ijf4eolrrx15) 3

[2.1.3 Select Entity Extraction](#_qnfoguksy5te) 4

[2.1.4 Select the Dataset](#_9yz0u4v3p6co) 4

[2.2 Step 2 - Tagging Tool: Landing Page](#_iae2nugwapn) 5

[2.3 Step 3 - Labeling Screen](#_t5i4zkigqb0n) 5

[2.3.1 Documents Tagging](#_9t6nypm3gs37) 6

[2.3.2 Choose Appropriate key](#_qzj1awf2cnb5) 7

[2.3.3 Delete Key Value](#_8r8c0u9xrtgj) 7

[2.4 Ground Truth Values For Untagged Documents](#_9eomvhxxto7g) 8

[2.4.1 Access to ‘lease-poc-quantiphi-untagged’ bucket](#_jv8c04rgduz) 8

[2.4.2 Access The PDF Files](#_errwowcaf2sg) 9

[2.4.3 List Down The Ground Truth Values/Page no.s](#_yap8rmoly7vq) 9

[**3. Appendix**](#_tr0oz051pitd) **10**

[3.1 Labels need to be tagged](#_v8tk3qwnhpbe) 10

[3.2 Training Plan](#_9zqqpawwtwhm) 11

[3.2.1 Agenda for Thursday (08/19/2021)](#_43fbzchwvjcw) 11

[3.2.2 Tagging Daily Workload Tracking](#_hsw6akfvs1t1) 11

[3.3 Instructions/POC](#_h2xggvslrs07) 11

[3.4 KT session recordings](#_33ktnal2h20l) 12

[3.5 Steps to be followed to get GCP access](#_9l5otnlf04zt) 12

# 1.Introduction

This document is a step by step guide for a tagger to label data for entity extraction. The objective of this exercise is to export ground truth data for accuracy calculations.

This is currently hosted on JLL’s GCP Project.

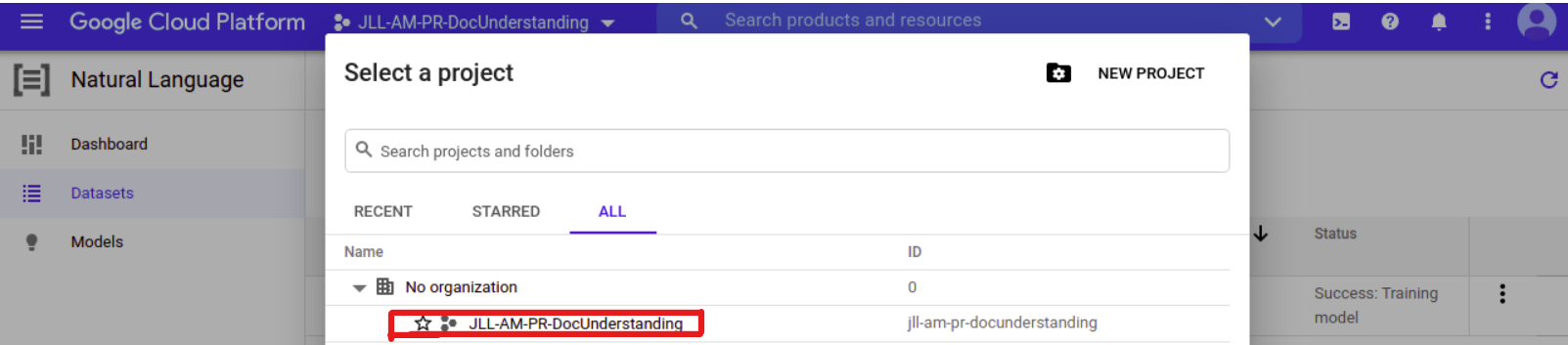
# 2.Procedure

## 2.1 Step 1 - Access to the GCP console

Users will click on this [link](https://console.cloud.google.com/home/dashboard?project=jll-am-pr-docunderstanding&authuser=0) to access the GCP Console

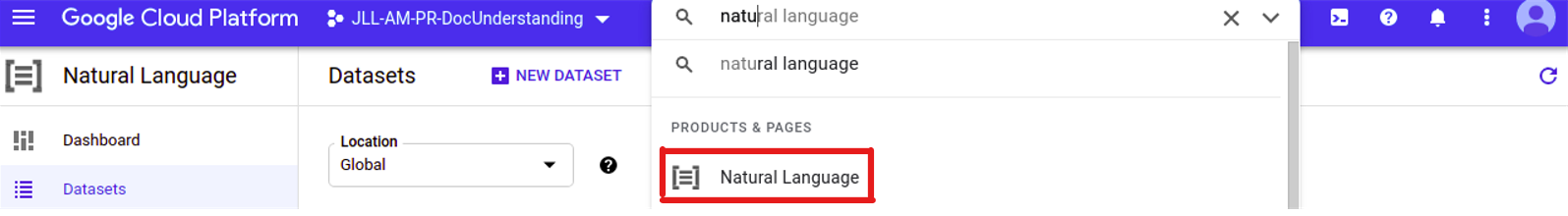
### 2.1.1 Select a project

Once the user enters into the GCP console, Select the Project: **JLL-AM-PR-DocUnderstanding** under **‘NO ORGANIZATION ’** as shown below in the image(Highlighted in red)



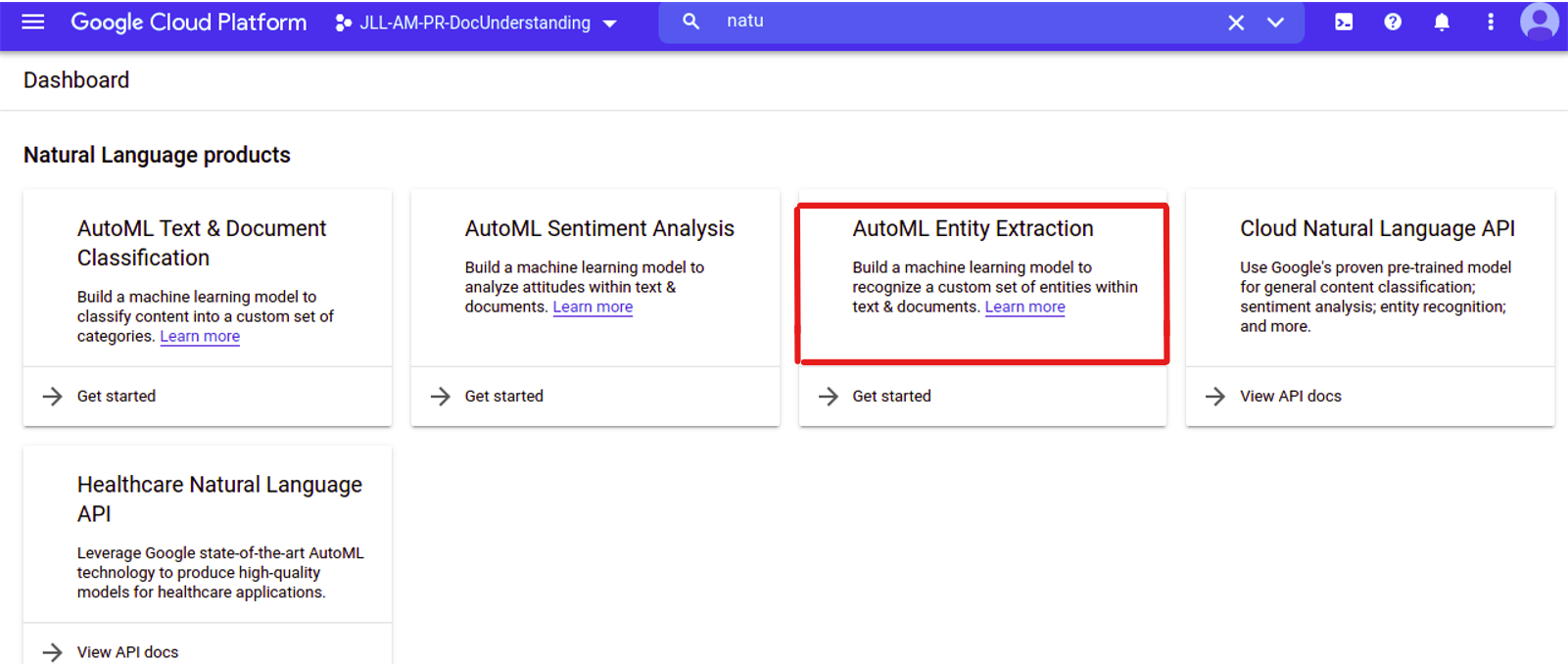
### 2.1.2 Select Natural Language Product

Once the user enters into the **JLL-AM-PR-DocUnderstanding** project, Type “**Natural Language**” in the search bar then select the “**Natural Language**” product as shown below in the image



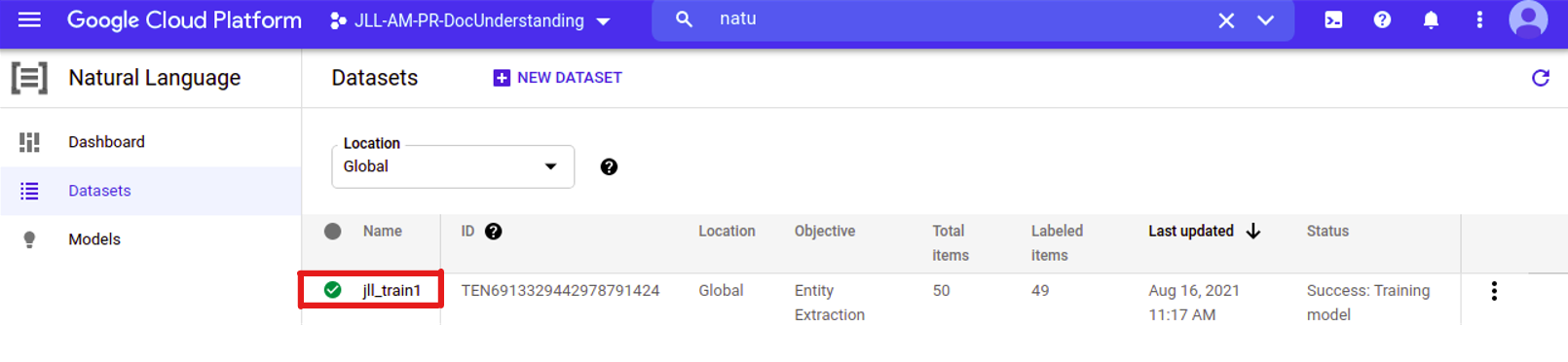
### 2.1.3 Select Entity Extraction

The user will route to all the natural language products as shown in the below image, select **AutoML Entity Extraction**

****

### 2.1.4 Select the Dataset

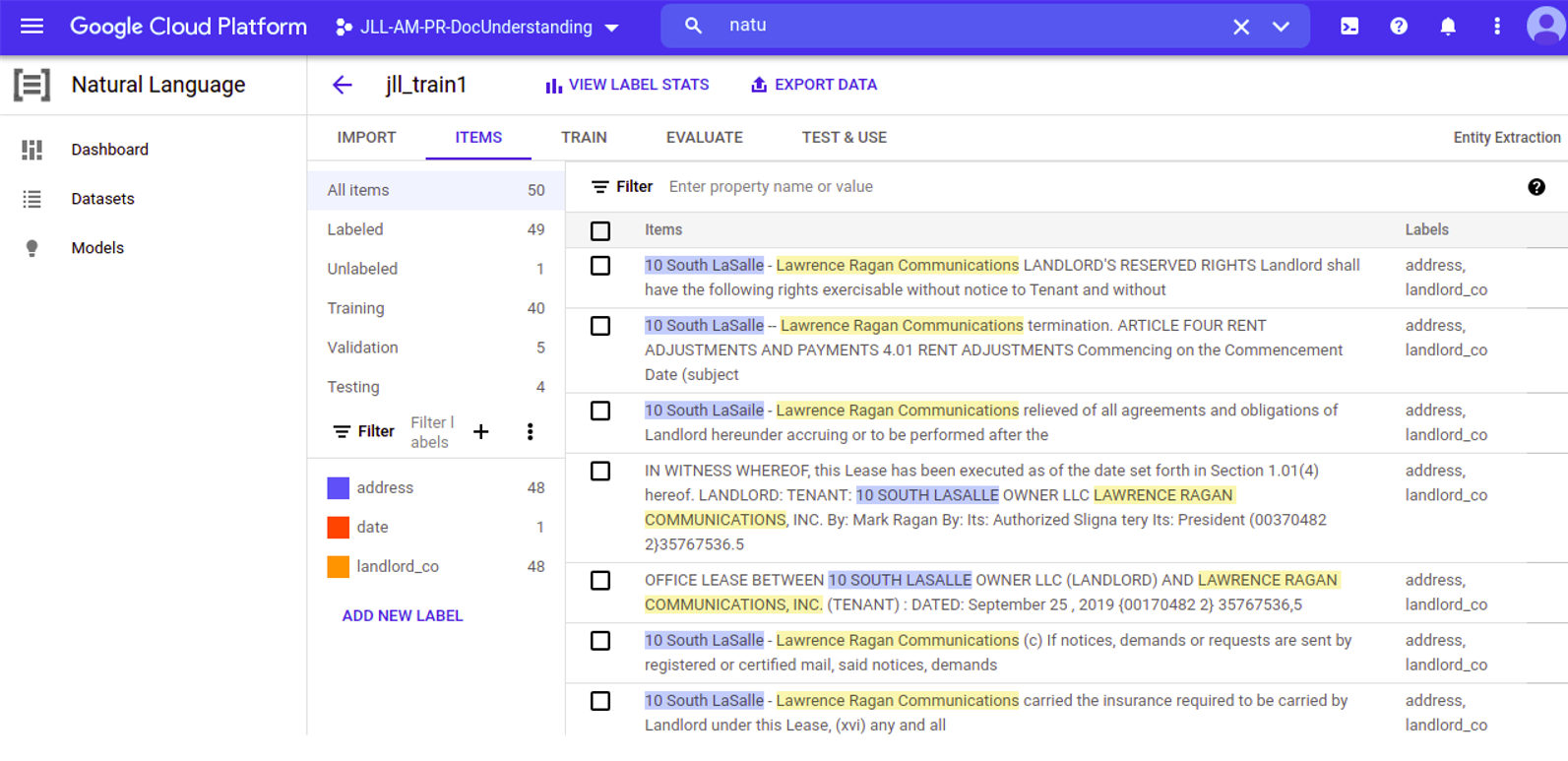
Now the user will be able to select a dataset (e.g., **jll\_train1** here) as shown in the below image



## 

## 2.2 Step 2 - Tagging Tool: Landing Page

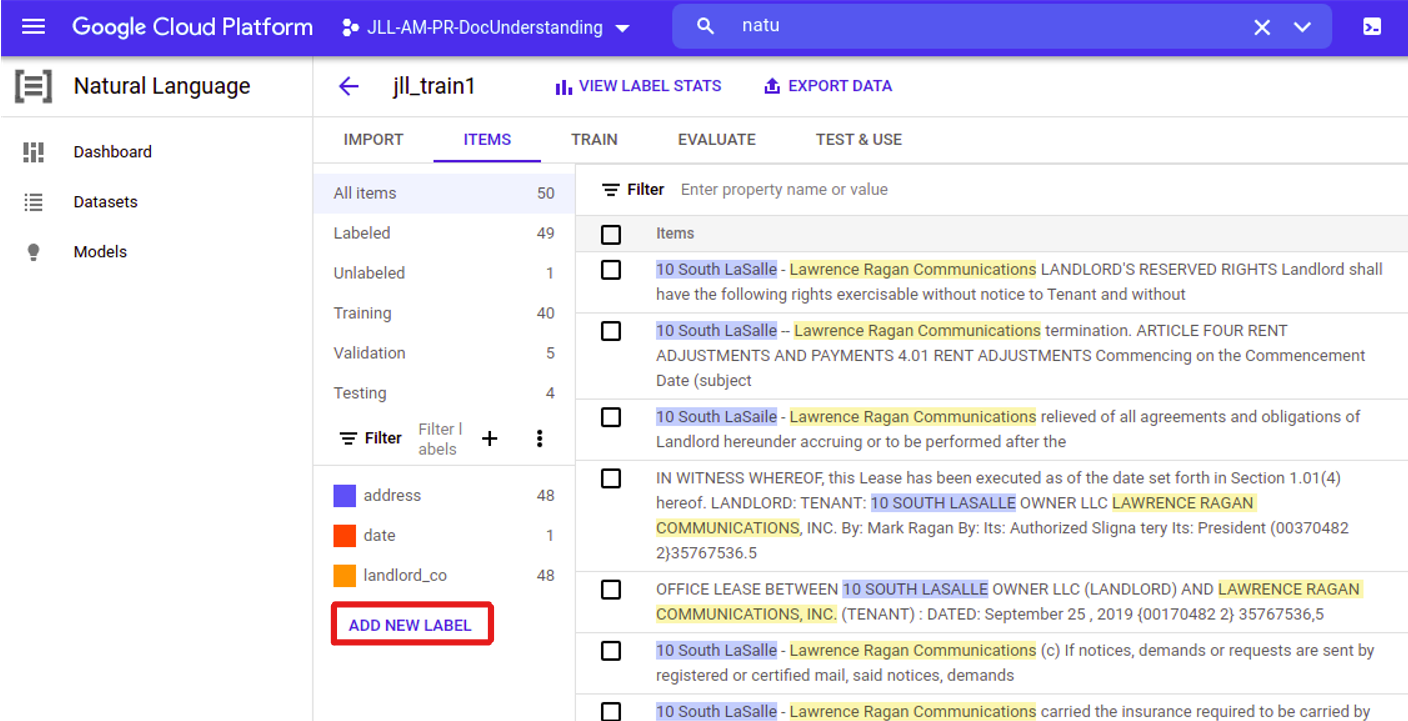
Once the user click on the dataset it will navigate to the documents present in that dataset



## 2.3 Step 3 - Labeling Screen

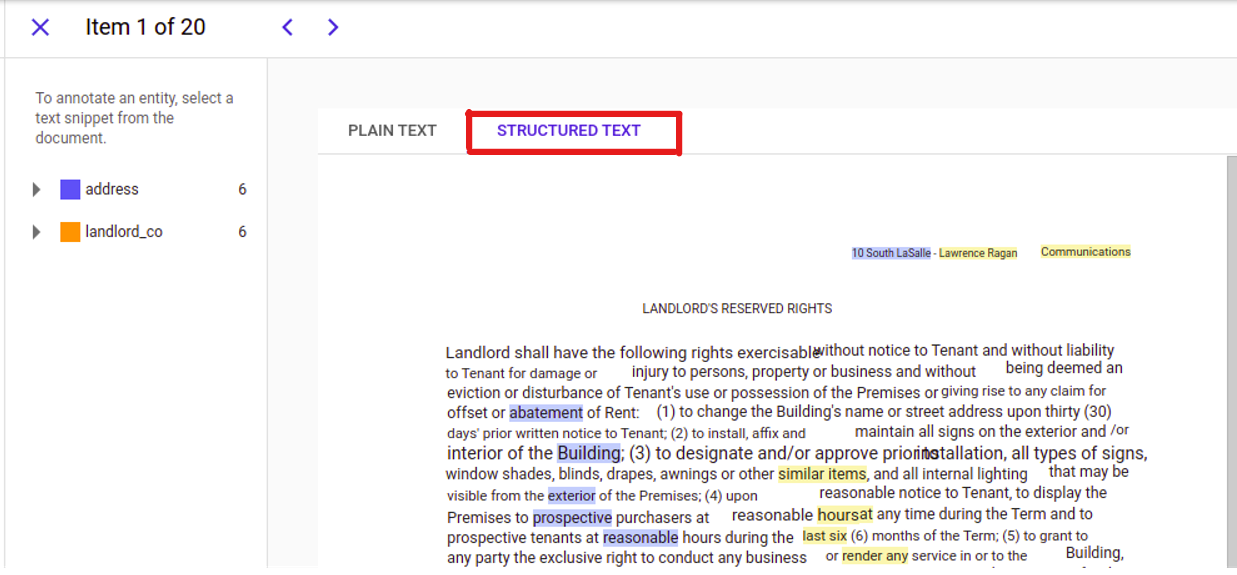
This screen will allow you to tag/label the data. On the left side, one can see the “**ADD NEW LABEL**” click on it to add a new label (field/key) & the right side is the actual document that is to be tagged. Original labels that are to be tagged are mentioned in the appendix section

**Note:** Users can add the entire list of labels prior to tagging that makes their work smoother/faster



### 2.3.1 Documents Tagging

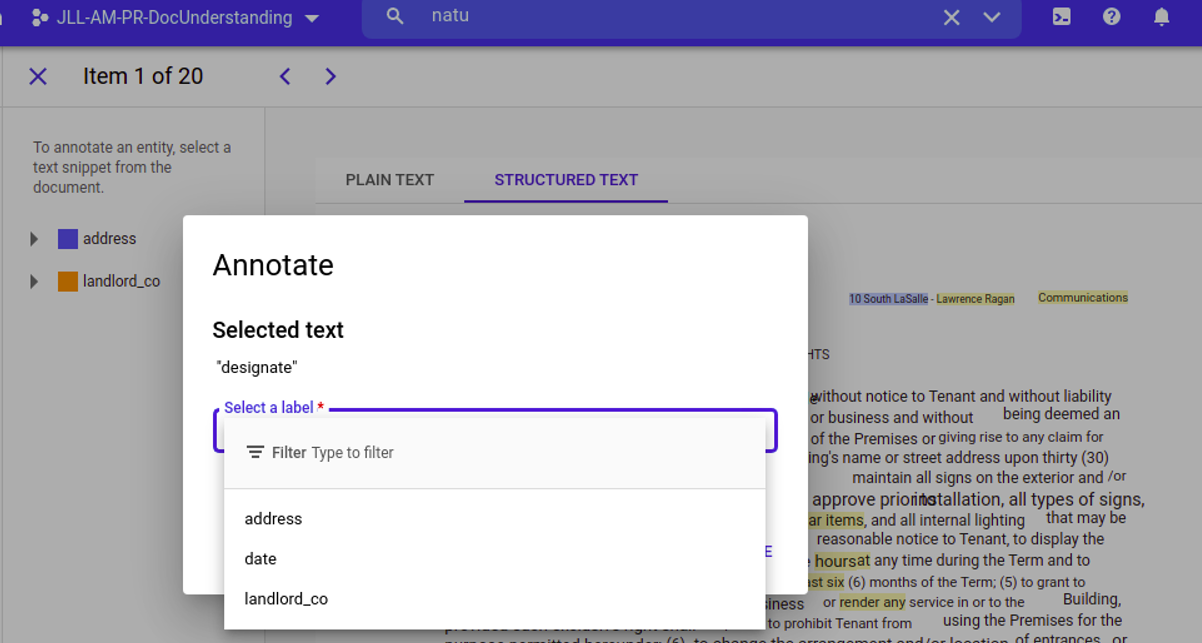
Double click on any of the documents to start tagging it will navigate the users to the “**STRUCTURED TEXT**” tab



**Note:** Do not tag the documents using “**PLAIN TEXT**” tab

### 2.3.2 Choose Appropriate key

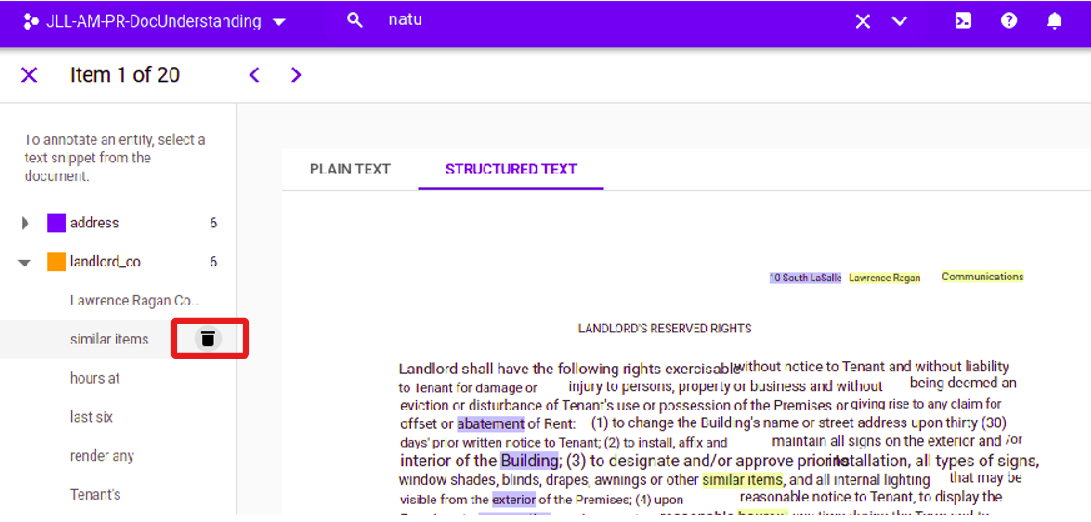
Select the appropriate text that users want to tag(highlighted in blue/yellow) then a drop down will appear with the list of labels, choose the appropriate field/key corresponding to the selected text/value



### 

### 2.3.3 Delete Key Value

In case of incorrectly tagged key value pair. Click on the delete icon as shown in the image below



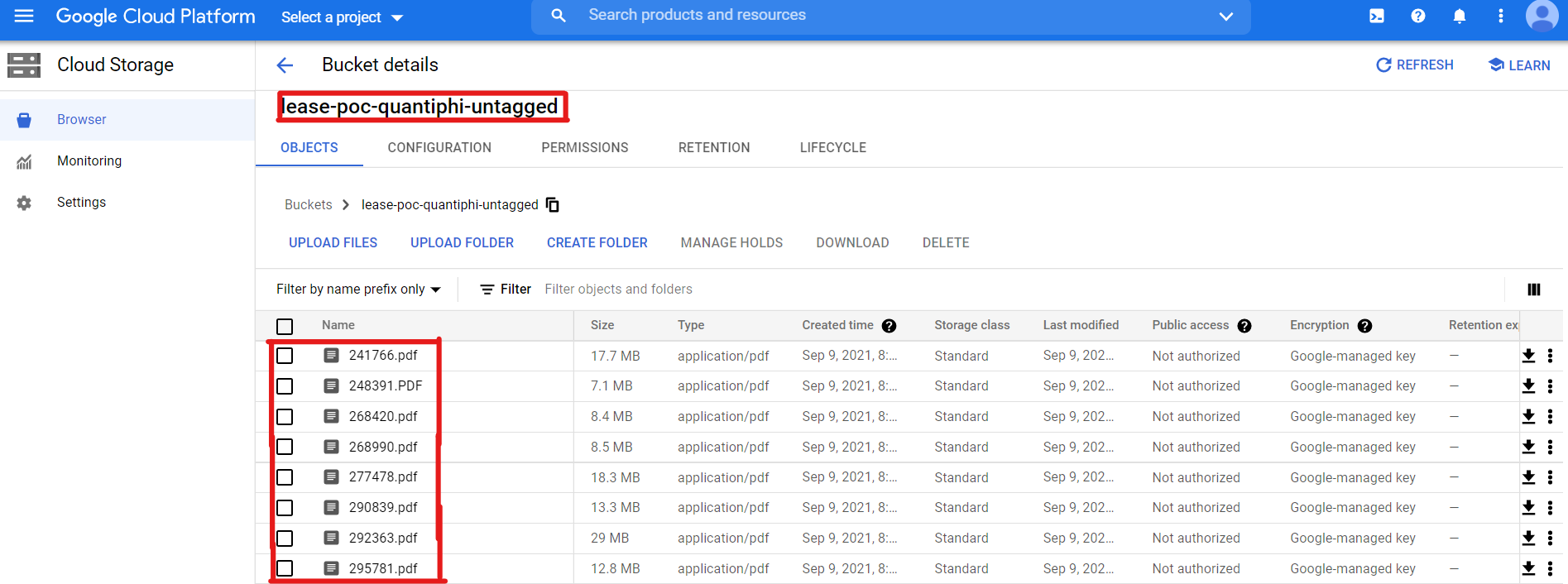
Repeat the above steps to retag the key values

## 2.4 Ground Truth Values For Untagged Documents

Users will work on exporting ground truth values along with the page numbers in the [spreadsheet](https://docs.google.com/spreadsheets/d/1KRcRl0q84V9CHXJs6pQipJPXvderXLuWnD1iCWqShwY/edit#gid=498742968) for validation purpose

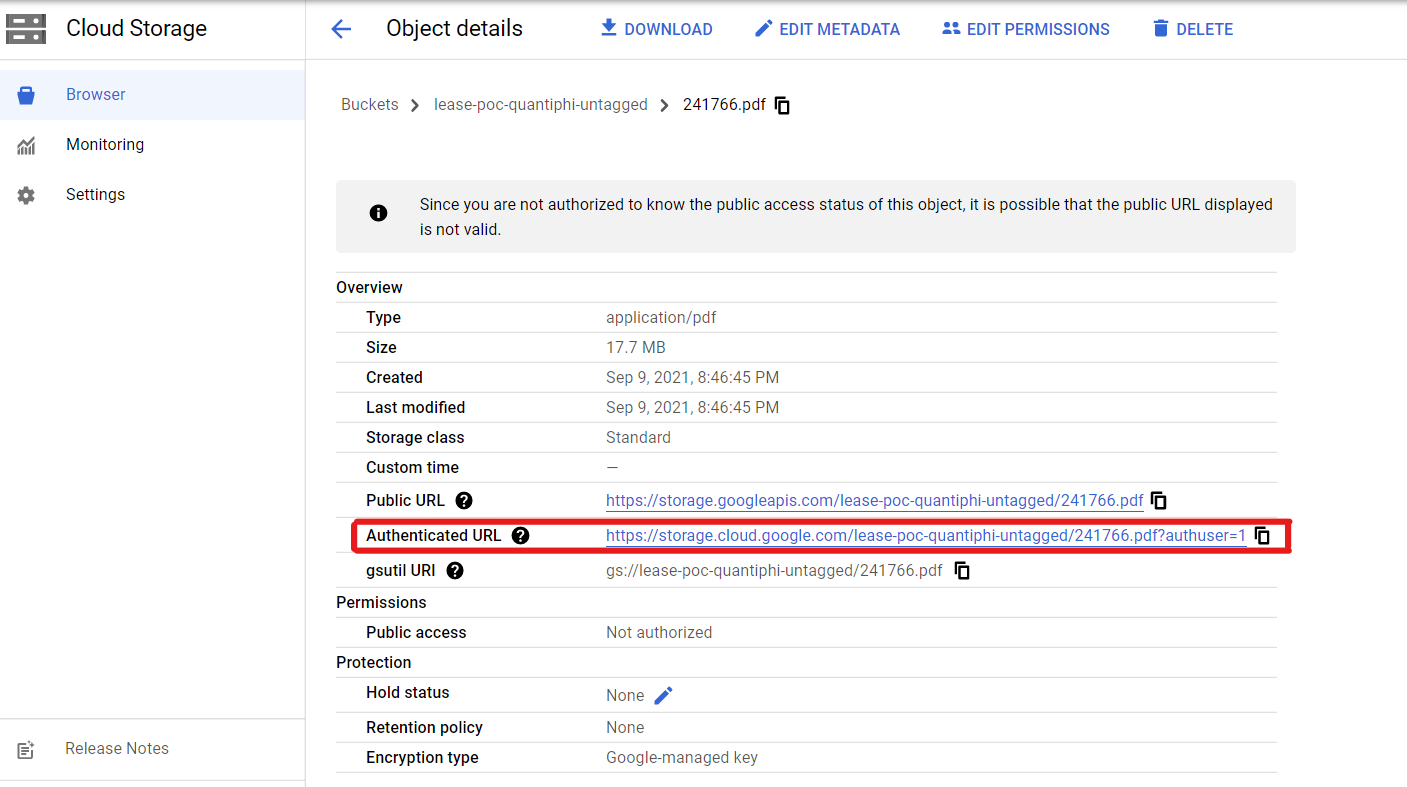
### 2.4.1 Access to ‘lease-poc-quantiphi-untagged’ bucket

Users will click on this [link](https://console.cloud.google.com/storage/browser/lease-poc-quantiphi-untagged) to access the untagged documents/pdf files as shown in the below image



### 2.4.2 Access The PDF Files

Once the user clicks on the pdf file it will navigate to an overview of the particular file as shown below in the image and then the user has to click on the **Authenticated URL** to view the pdf(by default it will open on a separate tab)



### 2.4.3 List Down The Ground Truth Values/Page no.s

Users will be able to view the entire pdf file then find all the entities and fill the values for each entity along with the page number in the shared [spreadsheet](https://docs.google.com/spreadsheets/d/1KRcRl0q84V9CHXJs6pQipJPXvderXLuWnD1iCWqShwY/edit#gid=498742968)

***Note:*** Users are not supposed to click on any of the download buttons while making ground truth values to untagged docs, it will trigger notification to the client and through a high level escalation, **Manish/Mahek** should alone take this responsibility to restrict the download of files

# 3. Appendix

## 3.1 Labels need to be tagged

1. Address\_Line\_1
2. Base\_Rent
3. City
4. Commencement\_Date
5. Expiration\_Date
6. Floor\_Number
7. Landlord\_Rep\_Company
8. Lease\_Size
9. Renewal\_Option
10. Renewal\_Option\_Notification\_Date
11. Security\_Deposit\_Amount
12. Sign\_Date
13. State
14. Suite
15. Tenant\_Company
16. Tenant\_Rep\_Company
17. Term\_Month
18. Zip\_Code
19. Renewal\_Option\_Date
20. Escalation\_Type
21. Base rent value

## 

## 3.2 Training Plan

* Monday (08/16/2021):
  + Onboarding - Introduction, project overview & document overview
  + Provision for GCP access
  + Walkthrough of the tagging tool (AutoML)
* Thursday (08/19/2021):
  + Training session1 - Detailed KT sessions
  + Define team structure (Taggers+Reviewer) & work shifts
  + Kickoff with a subset of data
* Friday (08/20/2021):
  + Training session2 - Review of tagged data, Q&A/Queries
  + Start with the actual dataset

### 3.2.1 Agenda for Thursday (08/19/2021)

1. Project Overview Deck - [Link](https://docs.google.com/presentation/u/0/d/1BKT9grbWhuuTiHOaPERNmSpH355bdigrmvwLJILYZSk/edit)
2. Document Overview - [Link](https://console.cloud.google.com/storage/browser/lease-poc-quantiphi/data/quantiphi-lease/batch1/pdf;tab=objects?pageState=(%22StorageObjectListTable%22:(%22f%22:%22%255B%255D%22))&authuser=1&project=jll-am-pr-docunderstanding&prefix=&forceOnObjectsSortingFiltering=false)
3. Define team structure/attendance tracker-[Link](https://docs.google.com/spreadsheets/d/1KRcRl0q84V9CHXJs6pQipJPXvderXLuWnD1iCWqShwY/edit#gid=649474478)
4. Detailed KT sessions over the documents - [Link](https://drive.google.com/drive/folders/1NNReu2F2DXWJ2b-8bj8UOo1H8BYrrO52)
5. Kickoff with a subset of data

### 3.2.2 Tagging Daily Workload Tracking

1. Enter details on taggers name, date, no of documents completed, and no of documents pending
2. In case if some file is corrupt or has some issues like not being readable or blurred mention in the generic comments tab

## 3.3 Instructions/POC

1. Please refer to this instructions manual for steps to follow while Tagging - [Link](https://docs.google.com/document/d/1b2vRD-imOEN0fFS1mOyZSFIDd-YTviq6UQ-V5NeL8UE/edit#)
2. In case of any queries, please post on the Google Chat - “**JLL: Tagging Team**” group or reach out to us via email  
   Satish Kumar - [satish.kumar@quantiphi.com](mailto:satish.kumar@quantiphi.com)

Naveen Shrivastava - [naveen.shrivastava@quantiphi.com](mailto:naveen.shrivastava@quantiphi.com)

Saicharan Gurramkonda - [venkata.gurramkonda@quantiphi.com](mailto:venkata.gurramkonda@quantiphi.com)

## 3.4 KT session recordings

1. KT Session 1- [link](https://drive.google.com/drive/folders/1NNReu2F2DXWJ2b-8bj8UOo1H8BYrrO52)
2. KT Session 2 - [link](https://drive.google.com/drive/folders/1NNReu2F2DXWJ2b-8bj8UOo1H8BYrrO52)
3. KT Session 3 - [link](https://drive.google.com/drive/folders/1NNReu2F2DXWJ2b-8bj8UOo1H8BYrrO52)
4. KT Session 4 - [link](https://drive.google.com/file/d/1Nx18FoTTzyoCOGHYldCRX1ZLnFdf-ZV9/view?usp=sharing) (Aug 26, 2021)
5. KT Session 5 - [link](https://drive.google.com/file/d/1QNIIp0LSRb_iTP5FwOiDqnlI8_khkxN8/view) (Untagged docs ground truth values/page no.s)

## 3.5 Steps to be followed to get GCP access

1. Download and install OKTA 2FA on your phone
2. Click on this link <https://jll.okta.com/> to set up 2FA
3. Enter your JLL id as above and click next
4. On the next page enter your id and the temporary password(Bibin Babu will be sent in a separate email) and click on submit
5. Next page configures OKTA 2FA and do the authentication

**Note:** For data tagging, taggers should only use JLL credentials to access the GCP