Document Digitization of Indian Passport

Business Case

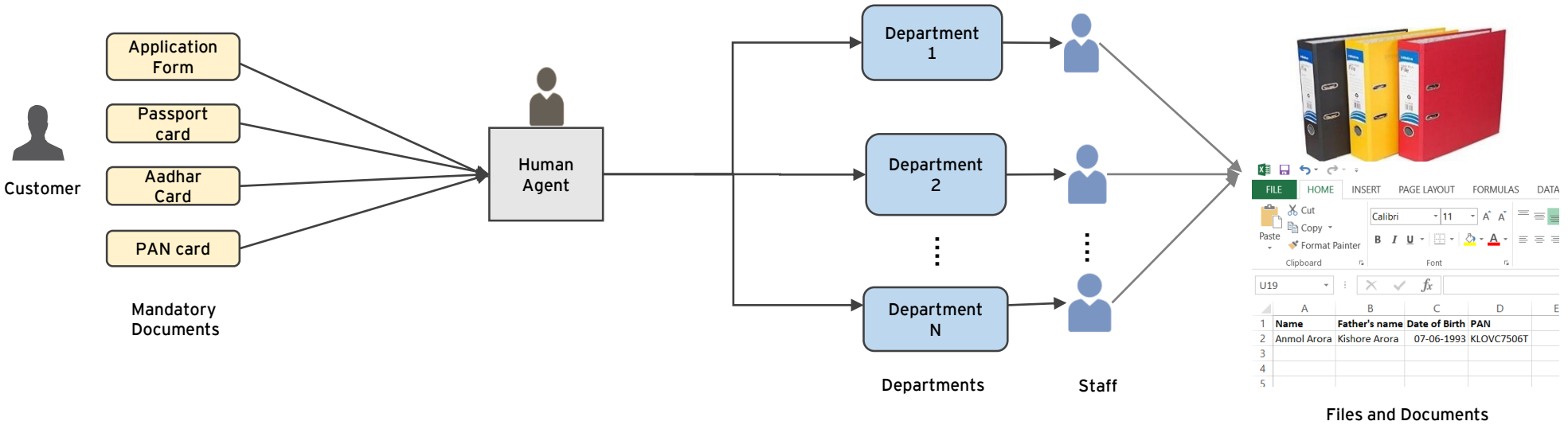
EY
Building a better working world

Problem Statement

Process of extracting data from Passport card and storing it manually in files is very time consuming task which requires lot of human effort.

Current Process / Solution

Customer submits documents → *Collecting and verifying documents* → *Submit to various departments* → *Store in System/files manually*



Challenges

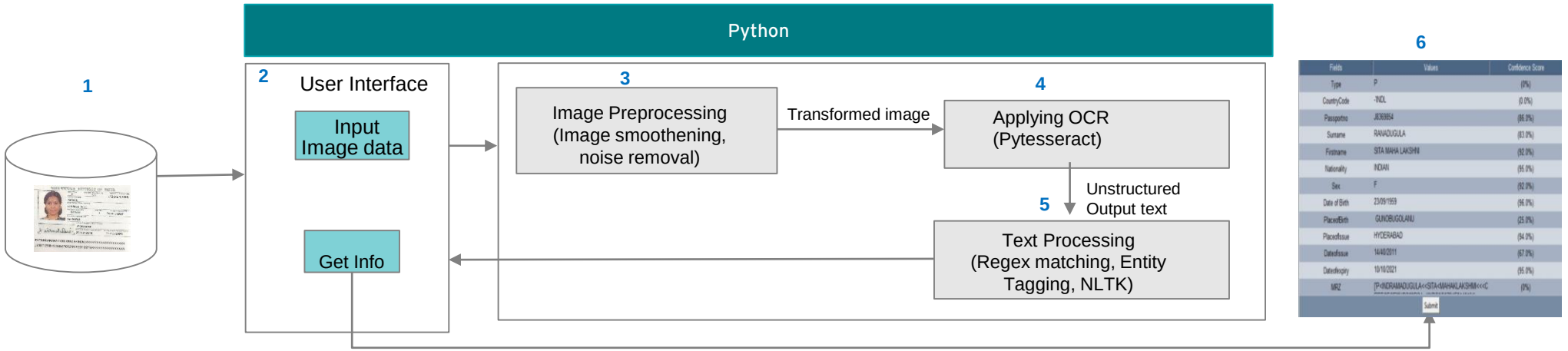
- Misplacement
- Disarrangement
- Searching
- Segregation
- Collaboration
- Document security problems

KYC of PAN card Approach

Objective: To develop and run an effective KYC application on Indian PAN card images, in order to extract data for in-person verification with least human intervention.

Solution

Image are stored in System database → *Get Image and get output information data* → *Image Preprocessed* → *Applying OCR and text Processed* → *Visualizing data in new image window*



Benefits

- Shorter processing time
- Protection from Fraud/Money laundering
- Document linking and mailing

1. Passport card images are stored in the system.

2. There is a User Interface created in Python that will get image upload on the UI application window on which further processing is done.

3. Image is pre-processed using defined packages and functions of python.

4. After Pre-processing, Pytesseract OCR is applied which generates unstructured output text

5. Text processing is performed like regular expression pattern matching, Entity tagging using various python libraries. Finally a JSON file is generated.

6. Information is displayed in separate window as we click on 'Process' button on Passport KYC UI

Prerequisites :

Programming Language

- Python 3.6

Built-in Packages

- **Pytesseract** - OCR module
- **Opencv** - To preprocess images
- **Regex** - To find specialised syntax
- **PIL** - To load image for UI
- **NLTK** - To do Natural Language Processing on the extracted text
- **Datetime** - To process dates
- **Tkinter** - To make the UI
- **OS** - To read, write, and delete files from the system
- **JSON** - To save the extracted data in json form

User-Defined Packages

- **Scrolling_area** - In order to provide a scrolling interface in the information window
- **Tabless** - In order to display the structured output data in a table

System Platform

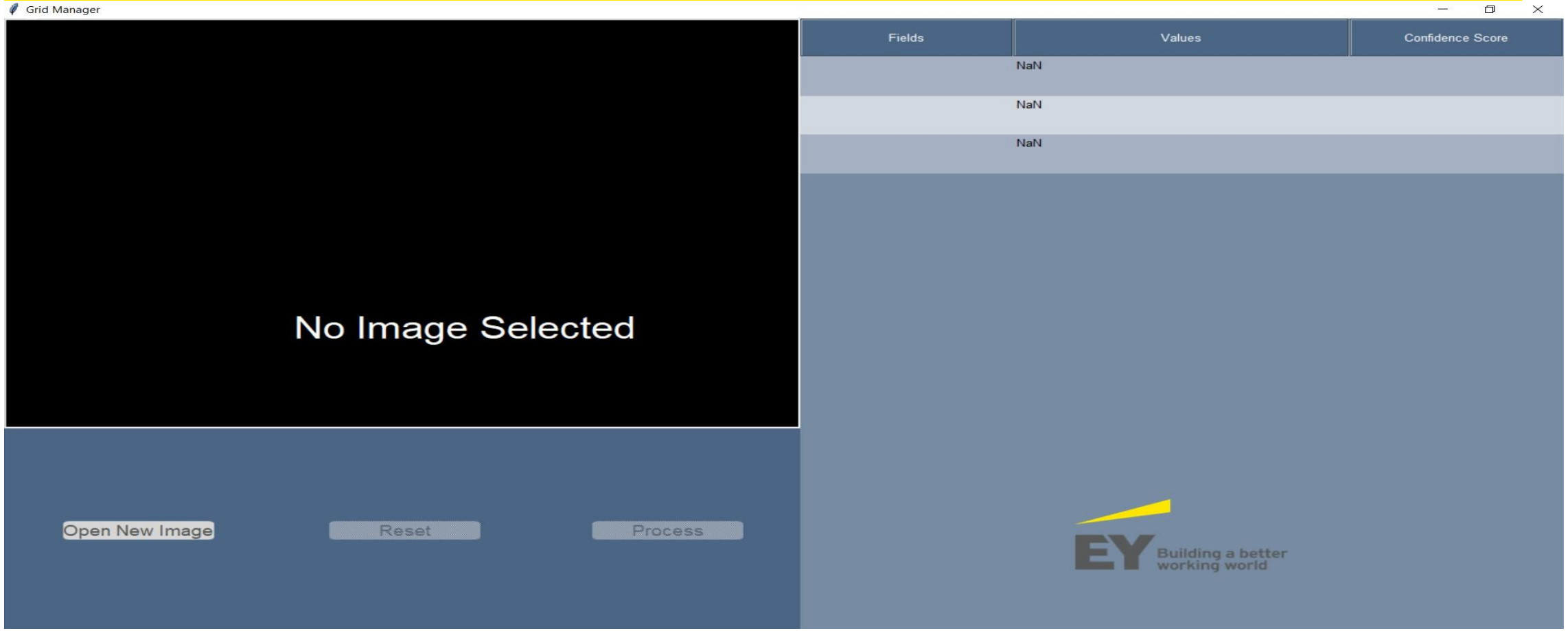
- Windows
- Mac
- Linux

Image Requirements

- Cropped Image of the Pan Card
- Size - Minimum 19 Kb

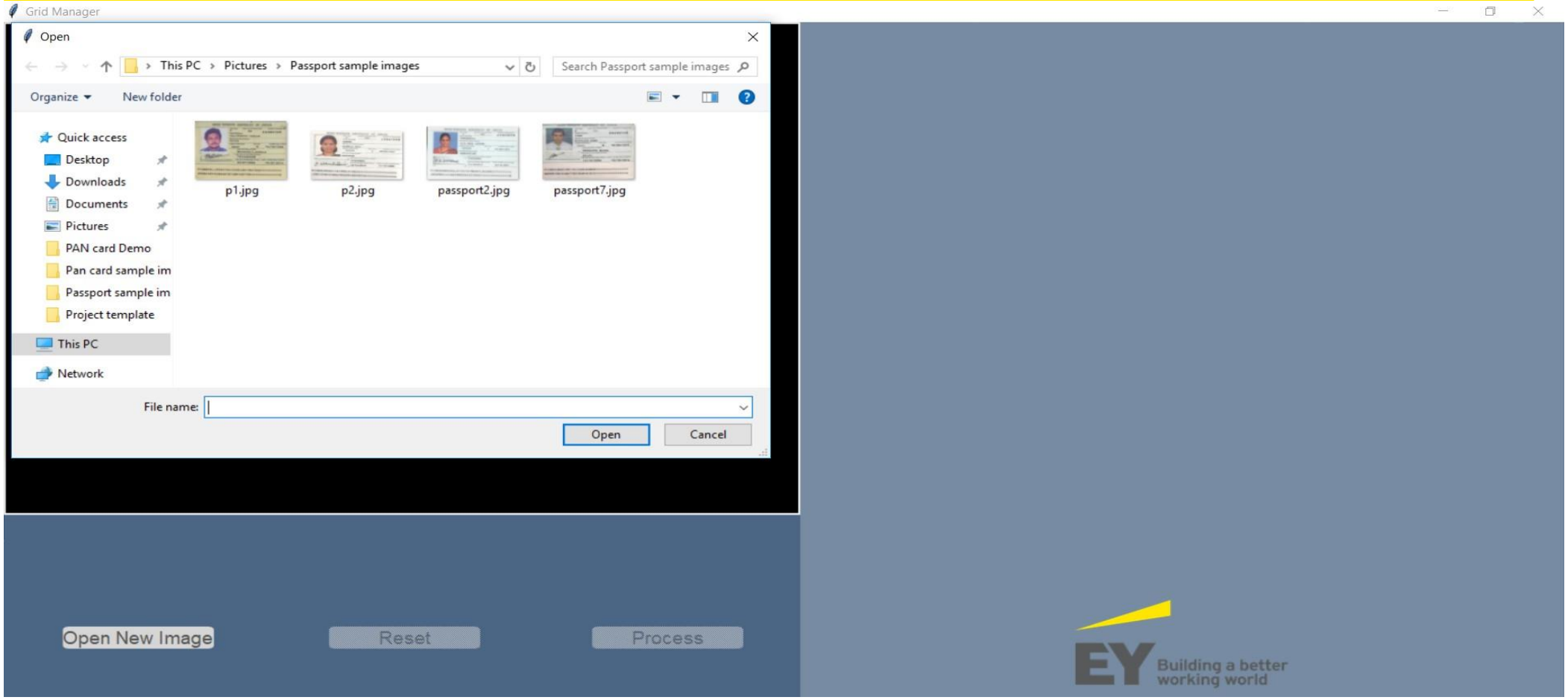
Application Execution:

Step 1: Passport Card Application Interface Window



Application Execution:

Step 2: Browse image and open it on Application window



Application Execution:

Step 3 : Image opens in Image viewer

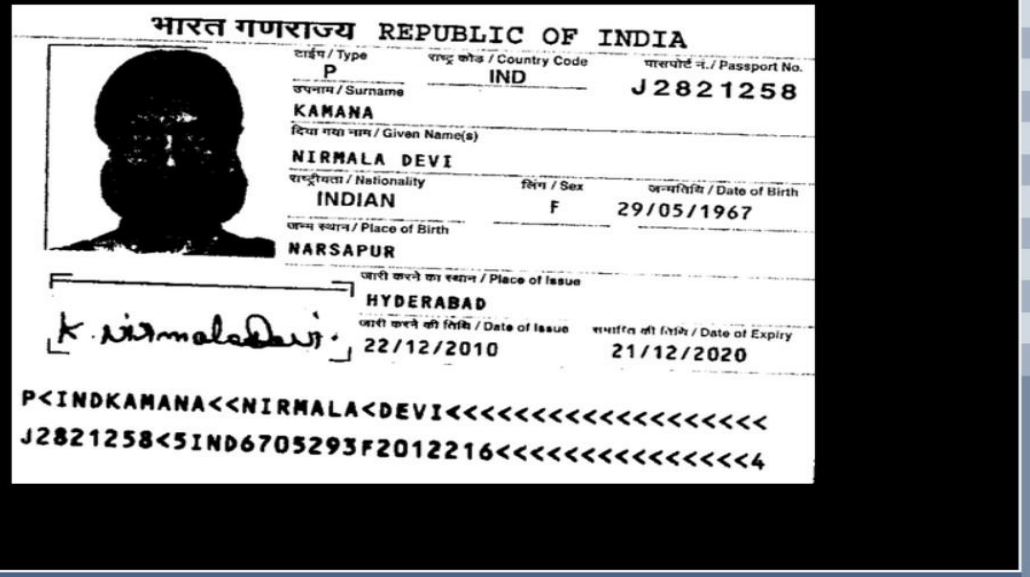
Application Execution:

Step 4 : Grayscale Image opens in Image viewer when process button is clicked

Application Execution:

Step 5 : Passport card Information is displayed on new window with confidence score

Grid Manager



Fields	Values	Confidence Score
Type	P	(83.0%)
CountryCode	NAN	(0%)
Passportno	J2821258	(90.0%)
Surname	KAMANA	(91.0%)
Firstname	NIRMALA DEVI	(81.5%)
Nationality	INDIAN	(92.0%)
Sex	F	(92.0%)
Date of Birth	29/05/1967	(95.0%)
PlaceofBirth	NARSAPUR	(91.0%)
Placeoflssue	HYDERABAD	(91.0%)
Dateofissue	22/12/2010	(95.0%)
Dateofexpiry	21/12/2020	(80.0%)
MRZ	['P<TNDKAMANAX<NIRMALACDEVI€<CCCCCCECEC	(0%)

Submit

Open New Image | Reset | Process

EY Building a better working world

Appendix A : Python Script for Passport card information extraction

UI2pass.py

Appendix B : Pssport card KYC media file