Environment and Climate Change Canada | Environnement et Changement climatique Canada

OPTICAL CHARACTER RECOGNITION OF CITES PERMITS USING MACHINE LEARNING

February 22-23, 2023

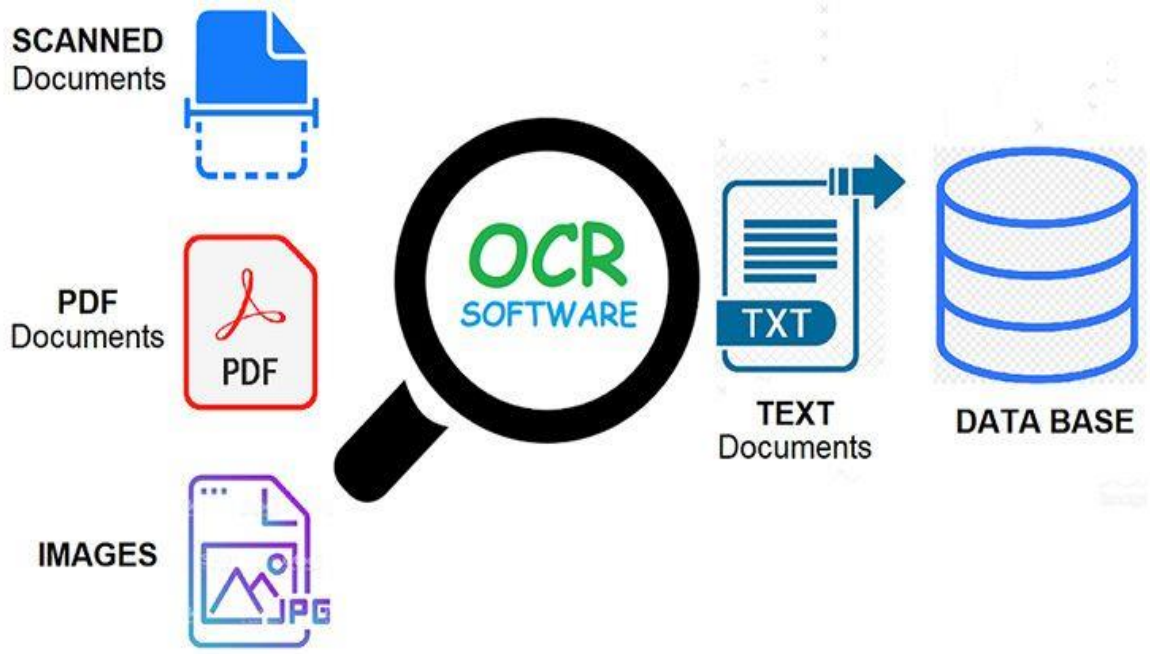
OFFICE OF THE CHIEF DATA OFFICER
STRATEGIC POLICY BRANCH

BY:
ARPIT RATHORE – Data Scientist
LAKSHAY GOEL – Software Developer (Co-op)
MICHAEL HOCKEY – Software Developer (Co-op)

WHAT IS OCR ?

- OCR stands for "**Optical Character Recognition**." It is a technology that recognizes text within a digital image.

- Optical character recognition (OCR) **converts printed paper documents into machine-readable text documents**.



WHAT ARE CITES PERMITS ?

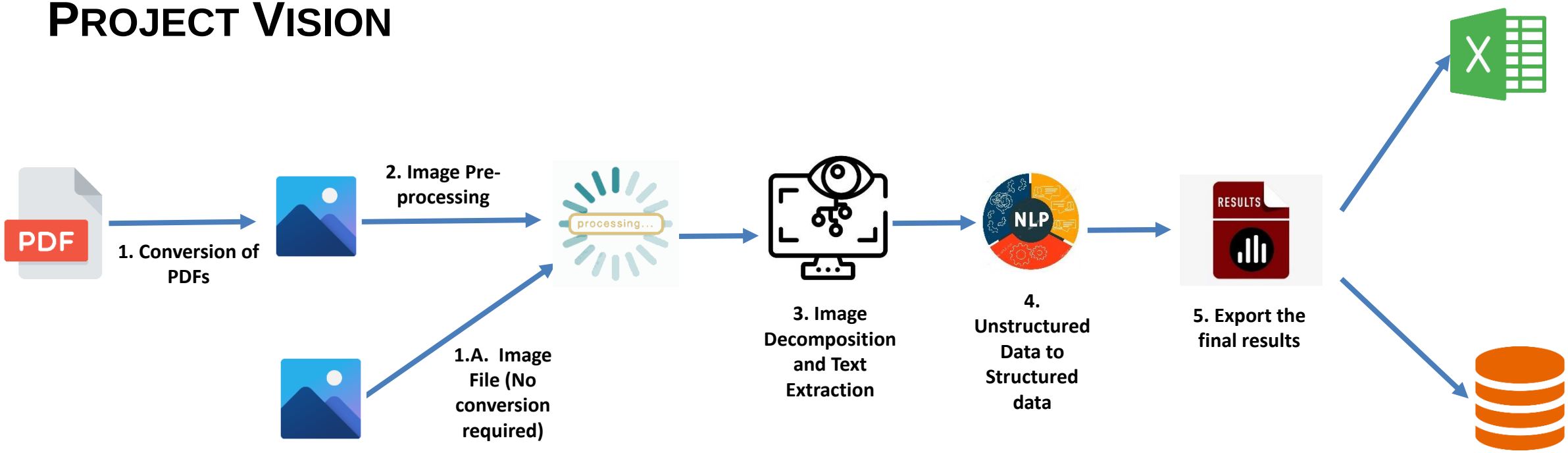
- The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) is **an agreement to ensure that international trade in wild animals and plants does not threaten their survival**.
- CITES permits are mandated for the importing or exporting commodities derived from CITES-protected species. This protection applies to the CITES-listed species in any form: alive or dead.
- The Canadian Wildlife Services (CWS) branch of Environment and Climate Change Canada oversees the administration of CITES Permits for Canada and other importing/exporting countries trading within Canada.

OFFICE OF THE CHIEF DATA OFFICER

PROJECT VISION

- Support the CWS branch on its ability to access the data within the CITES permits.
- Convert the thousands of scanned CITES permits from multi-year from various countries into a digital database using computer vision technology.
- Review and assess computer vision technology options (Azure Form Recognizer versus Python libraries) to select the optimal solution.
- Utilize the converted CITES permit information for further analytics.
- Provide a mechanism for CWS to upload and transform future permits into the database.

PROJECT VISION



PROJECT MILESTONES

1. **Options Analysis**
   - The Microsoft Azure Form Recognizer was tested to perform OCR but did not align with the nature of the CITES permits as it requires at least four samples of data for re-training purposes before extraction of data.
   - Python's pre-trained OCR models like PyTesseract and EasyOCR were tested but the information extraction was messy and could not be transformed into structured data.
   - Custom development in Python was therefore chosen as the optimal solution. The custom project flow and tool development incorporate Python libraries like OpenCV, EasyOCR, pillow, pdf2image, nltk etc..

2. **File Conversion**
   - Currently CITES permits are scanned and stored as multi-page PDF documents.
   - Created a data validation module to check the format of the incoming file.
   - Converted the PDF into images by sending it to data conversion module otherwise send it to the image pre-processing module.
   - The python packages mainly used for this task are PDF2Image, PIL etc.

PROJECT MILESTONES

3. **Image Pre-Processing**
   - Converted images are passed to image pre-processing module for creating uniform images, skewness correction, noise removal, RGB to Monochrome, Dilation etc. [SEE SLIDE 9, Image 1 to Image 2]
   - Image pre-processing help us to create a standardized data for template detection module.
   - The main python packages used for this task are OpenCV, Numpy etc.

4. **Template Detection, Image Decomposition and Text Extraction**
   - Pre-processed images will go through CITES permit template detection module which further help us during image decomposition where based on detected template the image decomposition module divide the image into n-number of smaller images. [SEE SLIDE 9, Image 3]
   - The decomposed n-number of images will go through OCR Text Extraction pipeline to scrape all text present within each small image. [SEE SLIDE 10, Image 4 to Image 5]
   - The main python packages used for this task are OpenCV, Pandas, PyTesseract, EasyOCR etc.

PROJECT MILESTONES

5 & 6. <u>Raw Data to Structured Data and Export Structured Data (In-Planning Phase)</u>

- Natural Language Processing (NLP) methodologies will be utilized for the conversion of raw data to structured data in order to deal with spelling errors and or inconsistent input.
- The python packages mainly used for this task are NLTK, Numpy, Pandas, Regex/Regular Expressions and NER etc.

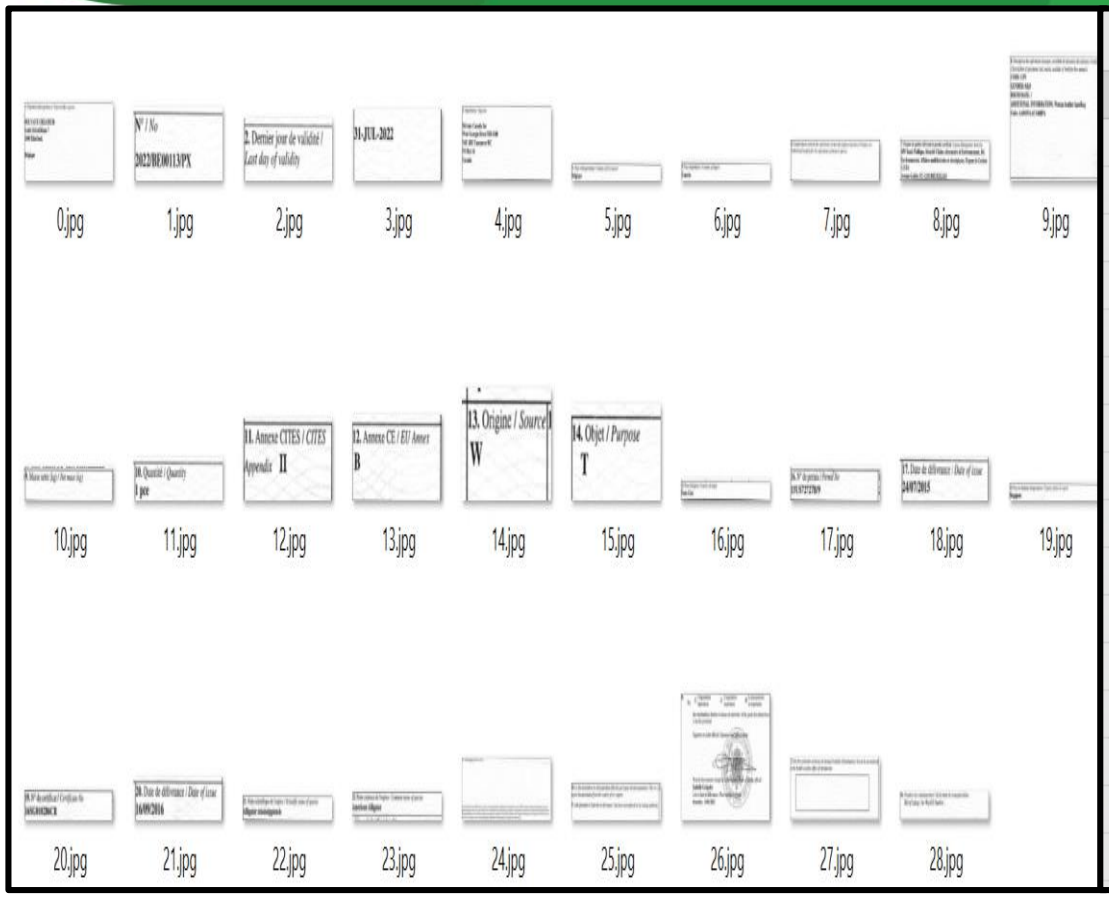
OFFICE OF THE CHIEF DATA OFFICER

1. Original Image

2. Pre-processed Image

3. Template Detection & Decomposition

OFFICE OF THE CHIEF DATA OFFICER



4. Cropped Images

	A	B	C	D	E	F	G	H	I	J
1	11. Export IDELVAUX [Belgique									
2	No / No									
3	9. Masse nette (kg) Net mass (kg)									
4	10. Quantité Quantity 1 pce									
5	11. Anncxc CITES CITES Appendix									
6	12. Anncxc CE EU Annex									
7	13. Origine Source									
8	14. Objct Purpose									
9	15. Pays dorigine 1 Country of origin États-Unis									
10	16. N' de permis Permii No									
11	17. Date de délivrance Date of issue									
12	18 Pays de demière réexportation / Country' of last re-export									
13	2. Dernier jour de validité Last day of validity									
14	19. No du certificat Certificate No									
15	22o. Date de délivrance Date of issue									
16	k1. Nom scientifique de Y'espècc Scientific name of species									
17	22. Nom commun de Tespèce / Common name of species									

5. Raw Extracted Data

QUESTIONS ??

Please reach out to me:

Arpit Rathore
Data Scientist
Office of the Chief Data Officer
Strategic Policy Branch, Environment and Climate Change Canada
Arpit.rathore@ec.gc.ca