Environment and Climate Change Canada | Environnement et Changement climatique Canada

RECONNAISSANCE OPTIQUE DE CARACTÈRES DANS LES PERMIS CITES À L'AIDE DE L'APPRENTISSAGE MACHINE

22-23 février 2023

BUREAU DU DIRIGEANT PRINCIPAL DES DONNÉES
DIRECTION GÉNÉRALE DE LA POLITIQUE STRATÉGIQUE

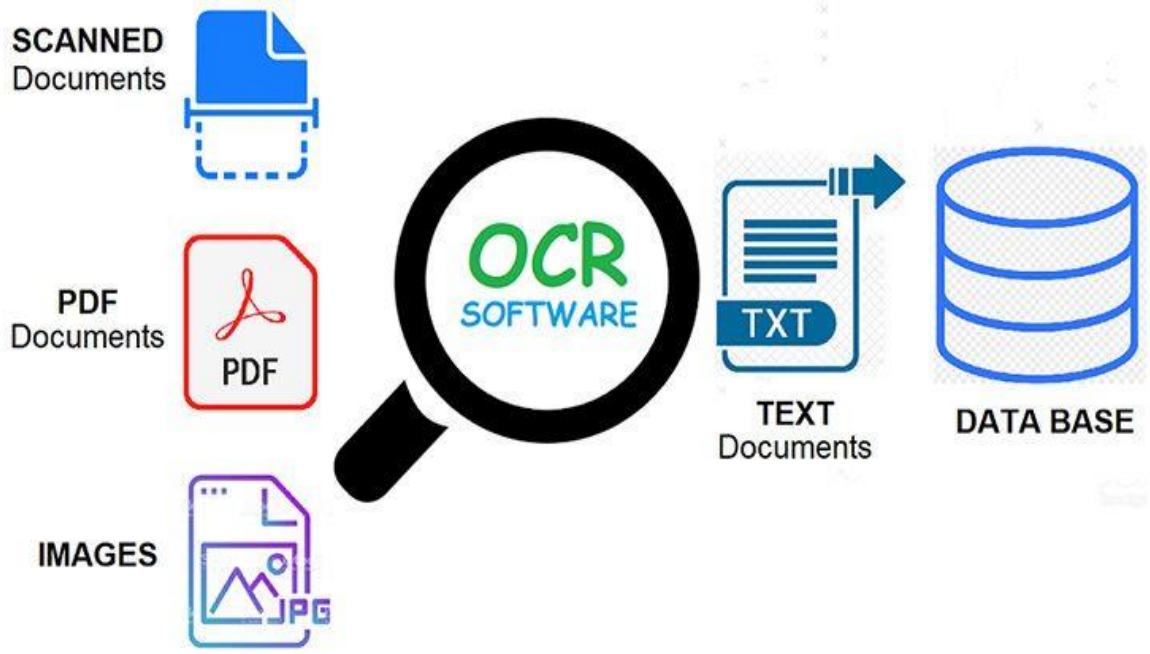
PAR :
ARPIT RATHORE – Scientifique des données
LAKSHAY GOEL – Développeur de logiciels (stagiaire)
MICHAEL HOCKEY – Développeur de logiciels (stagiaire)

Canada

QU'EST-CE QUE LA ROC?

- L'acronyme « ROC » désigne la **reconnaissance optique de caractères.** C'est une technologie qui reconnaît le texte dans une image numérique.

- La ROC **convertit les documents imprimés en documents-texte lisibles par une machine**.



QUE SONT LES PERMIS CITES?

- La Convention sur le commerce international des espèces de faune et de flore sauvages menacées d'extinction (CITES) est **une entente dont l'objectif est de veiller à ce que le commerce international des animaux et des plantes sauvages ne menace pas leur survie**.
- Un permis CITES est obligatoire pour l'importation et l'exportation de produits dérivés des espèces protégées par la CITES. Cette protection s'applique aux espèces inscrites à la CITES, quelle que soit la forme de l'individu (mort ou vivant).
- Le Service canadien de la faune (SCF) d'Environnement et Changement climatique Canada supervise l'administration des permis CITES pour le Canada et les autres pays importateurs ou exportateurs effectuant des échanges commerciaux au Canada.

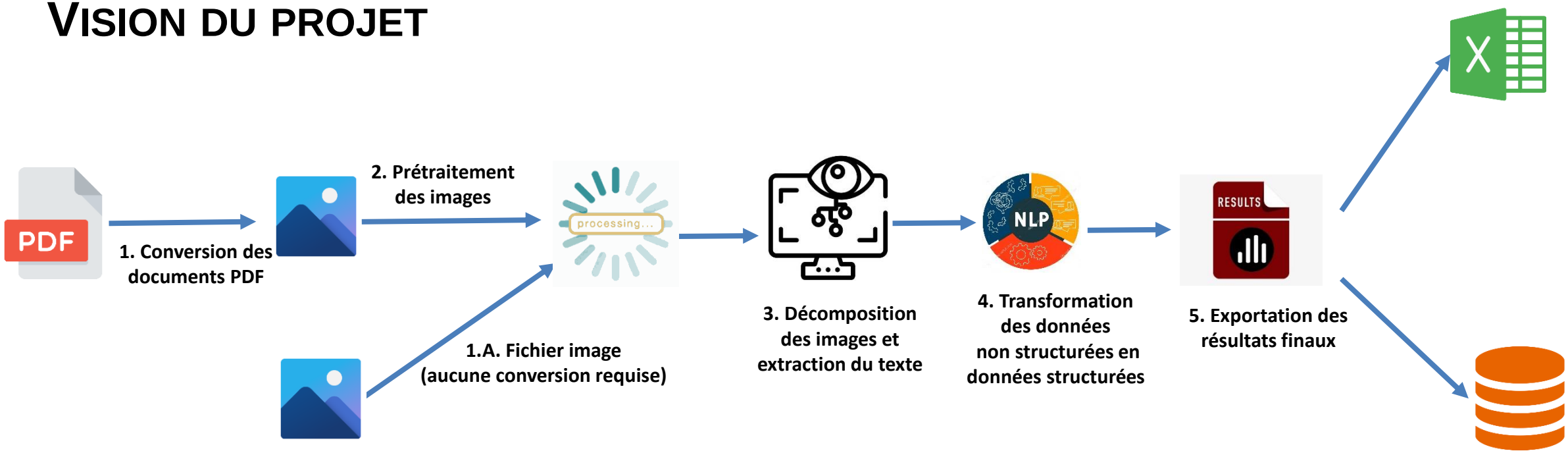
BUREAU DU DIRIGEANT PRINCIPAL DES DONNÉES

VISION DU PROJET

- Soutenir la capacité du SCF d'accéder aux données des permis CITES.
- Convertir, au moyen d'une technologie de vision informatique, les milliers de permis CITES numérisés de multiples années et de divers pays pour former une base de données numériques.
- Examiner et évaluer des options de technologie de vision informatique (Azure Form Recognizer et des bibliothèques Python) pour choisir la solution optimale.
- Utiliser les données converties des permis CITES à des fins d'analyse approfondie.
- Mettre au point un mécanisme pour que le SCF puisse téléverser de futurs permis dans la base de données en les transformant.

BUREAU DU DIRIGEANT PRINCIPAL DES DONNÉES

VISION DU PROJET



JALONS DU PROJET

1. Analyse des options

- L'outil Microsoft Azure Form Recognizer a été mis à l'essai pour la ROC, mais il ne convenait pas à la nature des permis CITES, car il a besoin d'au moins quatre échantillons de données pour procéder à un nouvel entraînement avant de passer à l'extraction.
- Des modèles Python préentraînés pour la ROC, comme PyTesseract et EasyOCR, ont été mis à l'essai, mais l'extraction était brouillonne et la transformation en données structurées était impossible.
- Il a donc été déterminé que la solution optimale était une conception sur mesure en Python. Le flux de projet et la conception des outils personnalisés intègrent des bibliothèques Python comme OpenCV, EasyOCR, Pillow, pdf2image, NLTK, etc.

2. Conversion des fichiers

- À l'heure actuelle, les permis CITES sont numérisés et conservés sous forme de documents PDF de plusieurs pages.
- Nous avons créé un module de validation des données pour vérifier le format du fichier entrant.
- Nous convertissons les documents PDF en images en les envoyant au module de conversion des données; sinon, les fichiers sont envoyés au module de prétraitement d'image.
- Les principaux ensembles Python servant à cette tâche sont pdf2image, PIL, etc.

VISION DU PROJET

3. Prétraitement d'image

- Les images converties passent par le module de prétraitement d'image, qui crée des images uniformes, corrige l'obliquité, élimine le bruit, convertit le RVB en monochrome, dilate les images, etc. [VOIR LA DIAPO 9, image 1 à image 2]
- Ce prétraitement aide à créer des données normalisées pour le module de détection de modèles.
- Les principaux ensembles Python servant à cette tâche sont OpenCV, NumPy, etc.

4. Détection de modèles, décomposition d'image et extraction de texte

- Les images prétraitées passent par le module de détection de modèles de permis CITES, qui facilite la décomposition des images : le module de décomposition se base sur les modèles détectés pour diviser l'image en *n* images de plus petites tailles. [VOIR LA DIAPO 9, image 3]
- Les *n* images décomposées subissent un processus d'extraction de texte par ROC, qui va chercher tout le texte présent dans chacune des petites images. [VOIR LA DIAPO 10, image 4 à image 5]
- Les principaux ensembles Python servant à cette tâche sont OpenCV, Pandas, PyTesseract, EasyOCR, etc.

VISION DU PROJET

5 et 6. <u>Transformation des données brutes en données structurées et exportation des données structurées (en cours de planification)</u>

- Des méthodes de traitement du langage naturel seront utilisées pour convertir les données brutes en données structurées, afin de gérer les fautes d'orthographe et les incohérences dans les données entrées.
- Les principaux ensembles Python servant à cette tâche sont NLTK, NumPy, Pandas, RegEx/Regular Expressions et NER, etc.

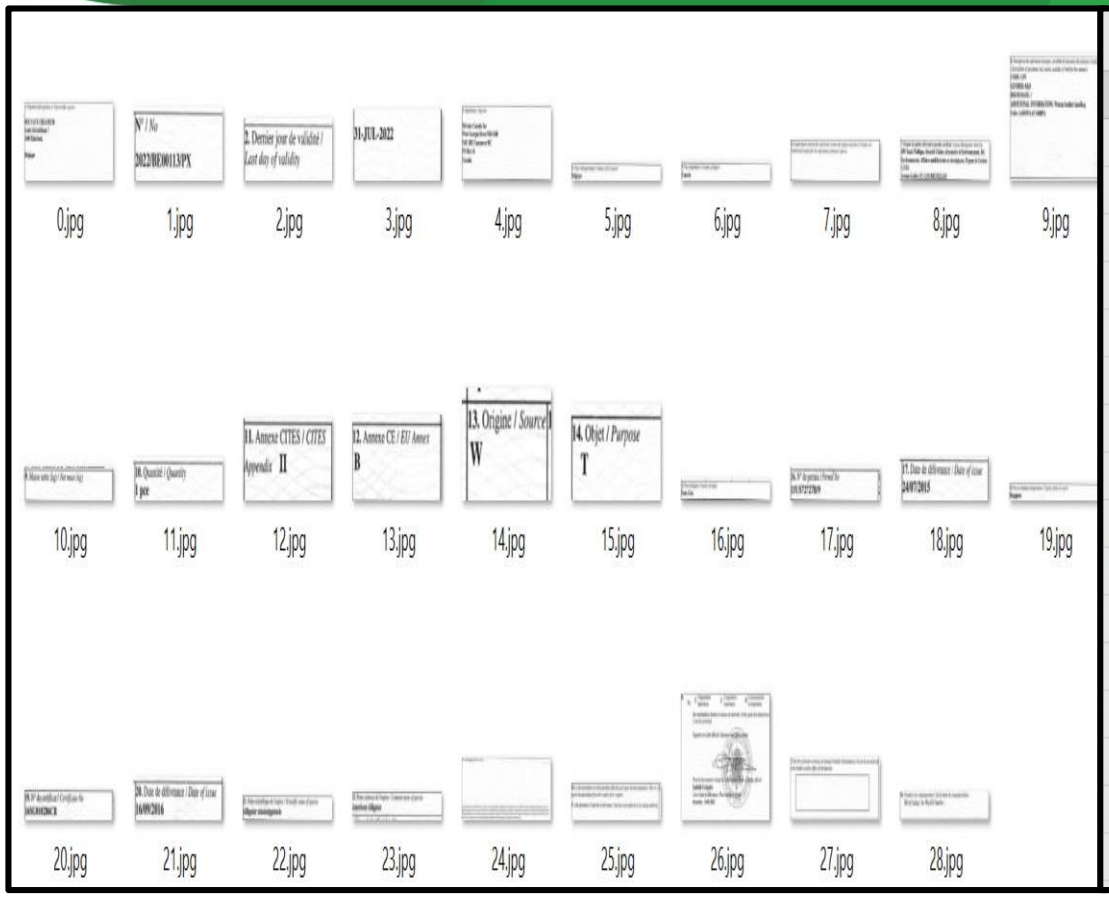
BUREAU DU DIRIGEANT PRINCIPAL DES DONNÉES

1. Image d'origine

2. Image prétraitée

3. Détection de modèles et décomposition

BUREAU DU DIRIGEANT PRINCIPAL DES DONNÉES



4. Images découpées

	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. Données extraites brutes

DES QUESTIONS?

Veuillez communiquer avec moi :

Arpit Rathore
Scientifique des données
Bureau du dirigeant principal des données
Direction générale de la politique stratégique, Environnement et Changement climatique Canada
Arpit.Rathore@ec.gc.ca