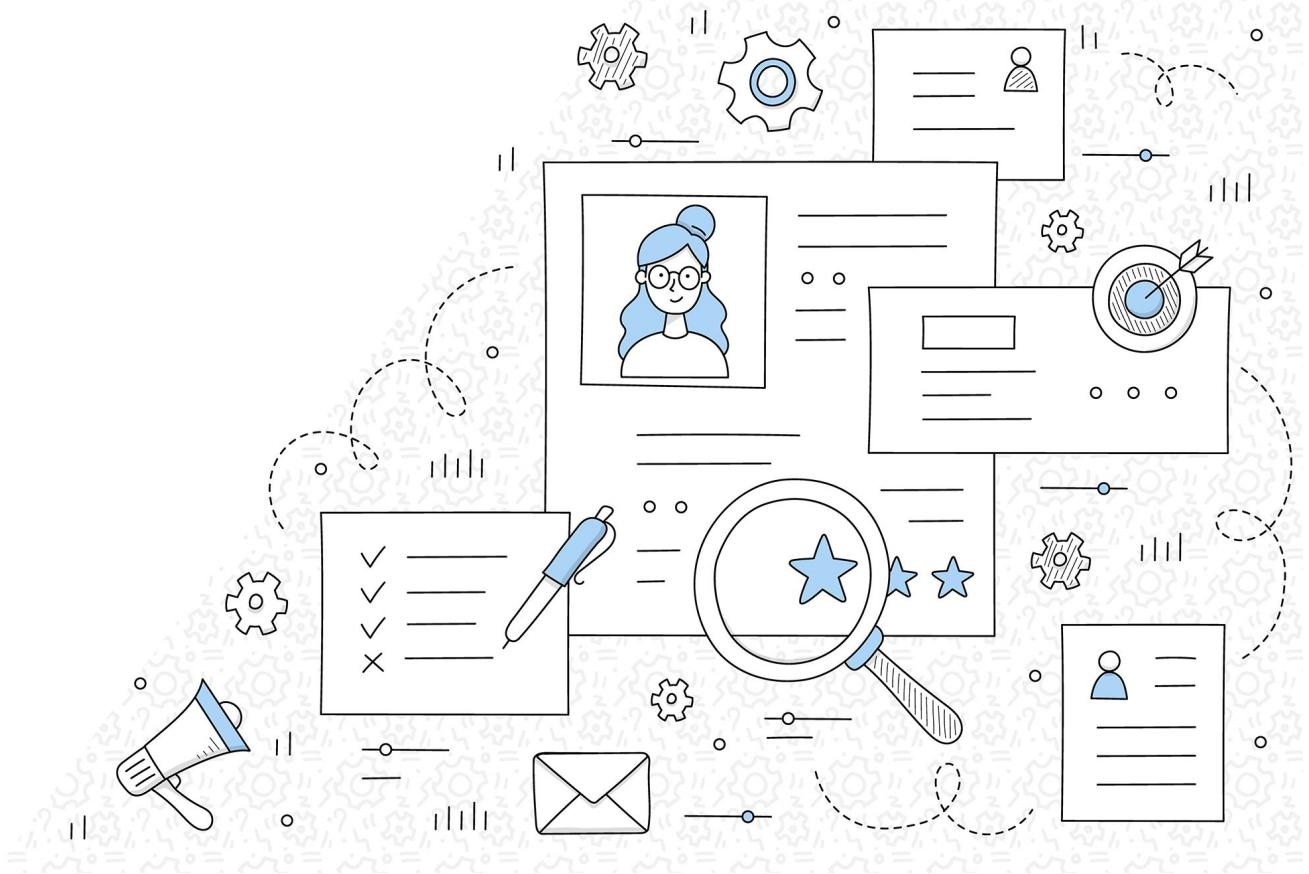
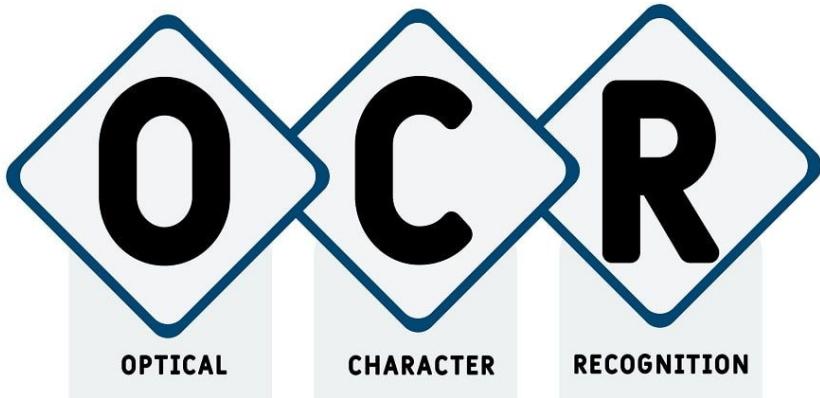


OCR

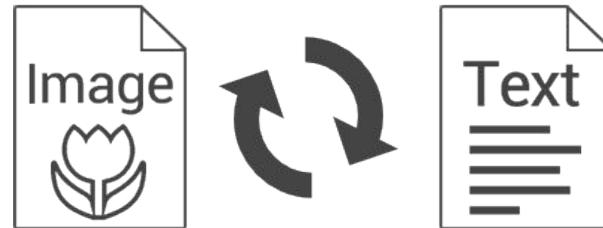


SOMMAIRE

1. Définition et application
2. Données et méthodes utilisées
3. Conclusion



La reconnaissance optique des caractères, ou *Optical Character Recognition* est une conversion électronique d'images textuelles dactylographiées, manuscrites ou imprimées. Ce texte est encodé pour l'obtenir en format texte.



Domaine d'application où l'OCR est le plus utilisée :

Domaine bancaire, assurances et valeurs mobilières

- Optimiser gestion documentaire d'une entreprise
- Détection de documents falsifiés
- Récupération d'information

DÉFINITION

Fonctionnement de l'OCR

L'algorithme de détection définit un caractère en évaluant ses lignes et ses traits.

La reconnaissance des motifs, fonctionne en identifiant l'ensemble du caractère.



Détection des lignes et traits



Reconnaissance des motifs sur un caractère



Reconnaissance des motifs sur une ligne de texte.

DÉFINITION

Fichiers format .pdf

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

Inspection and Surveillance Clearance Note				
Topic	Response	Comments	Notes	Responsibility
If inspection frequency target has been achieved for the batches(s) concerned?	Y	Except for activities that haven't occurred yet	Note: This relates to the completeness of the surveillance or the batch(es) at the time of release in the past.	Activity Lead/Region Leader
			(e.g. When PDR, it is important to indicate in the PDR if the inspection frequency target has been achieved. If not, a re-inspection meeting will be required (with M&E and QM) to propose the actions to be taken. A copy of the inspection report will be available with the issuance of the Clearance Note).	
If inspection target has not been achieved, provide evidence why this is acceptable for the batch(es) to be released.	N		For example, a provider of the goods that has not yet been implemented but have now been.	Activity Lead/Region Leader
			(e.g. Example: Provider of goods that are not yet implemented but have been tested with the quality of the products. On that guide test applied and not jeopardize the credibility of the supplier. In addition, the provider can indicate the date of the end of the post-ES surveillance (PMS) in comparison to the date of the inspection. This will give confidence on the quality of the product for the questions for which revised guides are issued).	Activity Lead/Region Leader
No. of Observation Sheets issued	N/A		N/A	Activity Lead/Region Leader
How many observation sheets have been satisfactorily responded to and closed at time of issuing the clearance note?	N/A		(e.g. Note all of your observation sheets. Consider closing any observations sheets relevant again.)	Activity Lead/Region Leader
Indicate if your inspection has been completed (grouped by GRC) or is expected to be completed by (date).	N		(e.g. If targeted, a final report will be required within 30 days of completion of inspection, in addition to a Surveillance Action Report.)	Engineering and Engineering Support Team Lead
			Important Note: Submission of the final inspection is required.	
Compliance Contact Details				
Role	Name	Date		
Quality Lead - AP (for E&S)	S. PRUD	Feb 2022		
Region Leader - NL (for E&S)		2022		
Manufacturing Engineer - ES (for E&S)				
Inspection Lead - IP (for E&S)				

- **Tabula**: Outil utilisé pour la lecture de tableaux de données dans les PDFs. Tabula est très utilisé par les sociétés d'informations.

The New York Times

PROPUBLICA

THE TIMES

- **pdfPlumber**: Package open-source de Python. Outil simple d'utilisation, et très efficace pour la récupération de texte et données dans les tableaux d'un PDF.
- **Tesseract**: Logiciel de reconnaissance optique de caractère. Peut lire des caractères de type UTF-8 dans plus de 100 langues.
- **OpenCV**: Bibliothèque mettant à disposition de nombreuses fonctionnalités, très connu pour le traitement d'image grâce aux opérations qu'elle permet d'effectuer.

LES MÉTHODES / LIBRAIRIES UTILISÉES

Fresh Bakery
3494 Rosewood Court
ELM MOTT
Texas

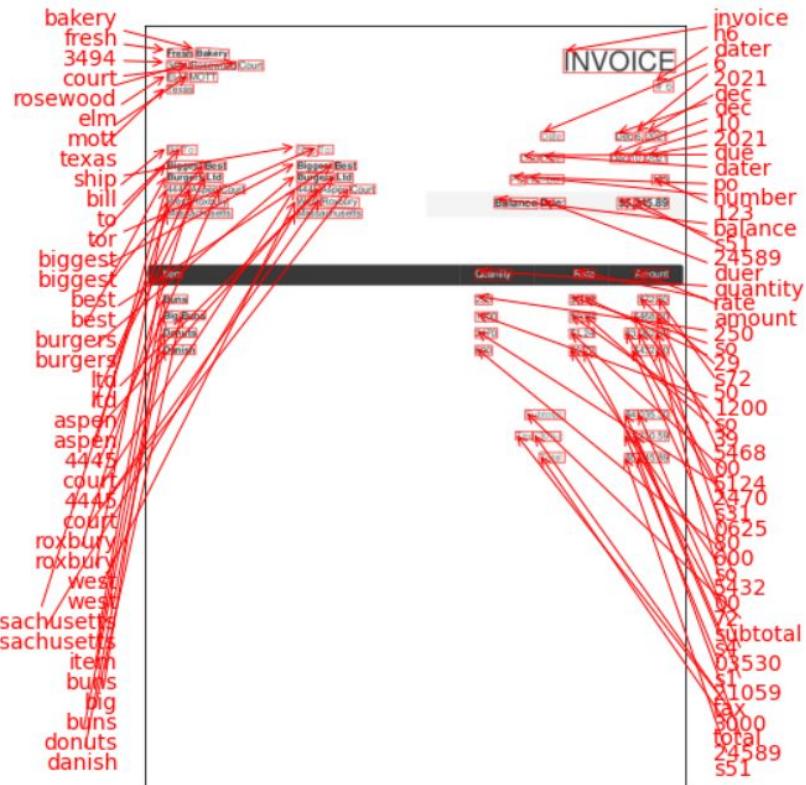
INVOICE

6

Bill To: Biggest Best Burgers Ltd 4445 Aspen Court West Roxbury Massachusetts	Ship To: Biggest Best Burgers Ltd 4445 Aspen Court West Roxbury Massachusetts	Date: Dec 6, 2021
		Due Date: Dec 10, 2021
		PO Number: 123
		Balance Due: \$5,245.89

Item	Quantity	Rate	Amount
Buns	250	\$0.29	\$72.50
Big Buns	1200	\$0.39	\$468.00
Donuts	2470	\$1.24	\$3,062.80
Danish	600	\$0.72	\$432.00

Subtotal: \$4,035.30
Tax (30%): \$1,210.59
Total: \$5,245.89



KERAS_OCR

Fresh Bakery
3494 Rosewood Court
ELM MOTT
Texas

INVOICE

6

Bill To:	Ship To:	Date:	Dec 6, 2021
Biggest Best Burgers Ltd 4445 Aspen Court West Roxbury Massachusetts	Biggest Best Burgers Ltd 4445 Aspen Court West Roxbury Massachusetts	Due Date:	Dec 10, 2021
		PO Number:	123
		Balance Due:	\$5,245.89

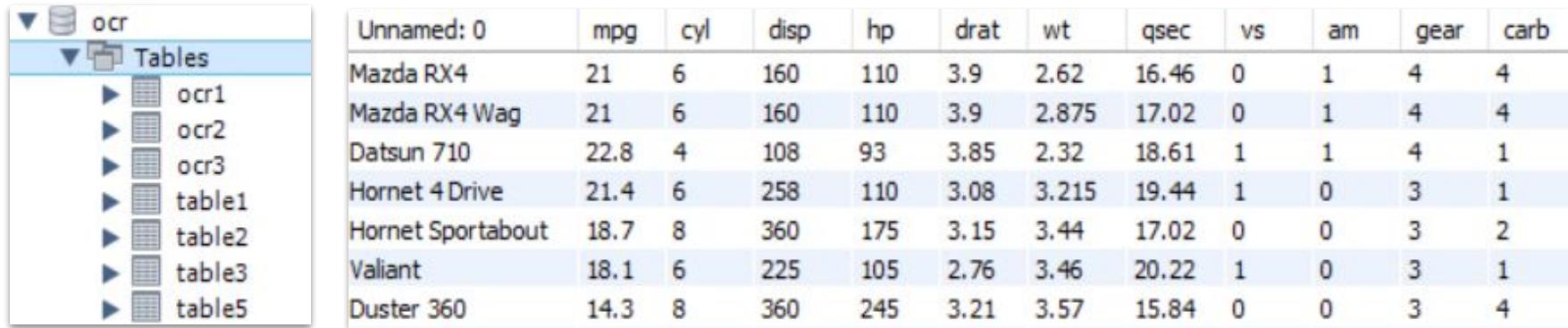
Item	Quantity	Rate	Amount
Buns	250	\$0.29	\$72.50
Big Buns	1200	\$0.39	\$468.00
Donuts	2470	\$1.24	\$3,062.80
Danish	600	\$0.72	\$432.00

Subtotal:	\$4,035.30
Tax (30%):	\$1,210.59
Total:	\$5,245.89



TESSERACT GOOGLE

- Export des données dans une base de données relationnelle via Python avec la librairie SQL Alchemy



	Unnamed: 0	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
▶	Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
▶	Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
▶	Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
▶	Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
▶	Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
▶	Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
▶	Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4

- Pour pallier aux erreurs d'exportation vers MYSQL notamment à cause de titre trop long dans nos tableaux, nous avons effectué une modification des données en Python.

MySQL

[Tesseract Google](#)

[OpenCV](#)

[PDFPLumber](#)

[Tabula](#)

[Comprendre l'OCR, le guide](#)



SOURCES