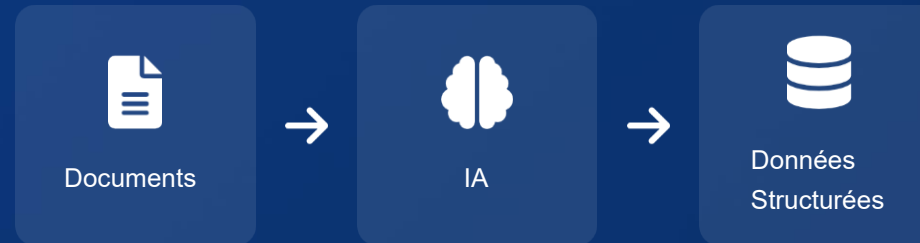


Système Intelligent d'Extraction et de Structuration de Documents RH par IA

Projet de Fin de Stage d'initiation - ONEE Branche
Électricité



 **Étudiant:**Hamza Elmourabit

 **Période de stage:**Juillet 2025

Soutenu le 20/09/2025 devant:

Mr.Lahcen

Moumoun

Mr.Nafidi Ahmed

 **Encadrement:**

- Mr. Mounir Chaïabi (Professionnel)

- Mr. Lahcen Moumoun (Académique)

Merci de votre attention et de votre soutien tout au long de ce projet



ONEE

Merci pour l'opportunité de réaliser ce stage enrichissant et pour votre accueil chaleureux



Encadrants

Merci à Mr. Mounir Chaïabi pour son encadrement professionnel et à Mr. Lahcen Moumoun pour son suivi académique




Jury

Merci à Mr. Lahcen Moumoun et Mr. Nafidi Ahmed pour votre temps et votre évaluation


"Ce projet a été réalisé avec beaucoup de motivation et de sérieux"

Introduction

 Page de titre

 Remerciements

Analyse du Projet


 Contexte et problématique

 Objectifs du stage


Approche Technique


 Technologies utilisées

 Méthodologie de travail

 Architecture du système

Implémentation

 Implémentation technique (1/2)

 Implémentation technique (2/2)


 Interface utilisateur

Résultats et Analyse


 Résultats obtenus


 Difficultés rencontrées

Future et Conclusion

 Perspectives d'évolution

 Apports du stage

 Conclusion

 Démonstration

Contexte et problématique



Contexte

L'ONEE, à l'instar de nombreuses grandes entreprises, s'engage activement dans la digitalisation de ses processus métiers. Cette transformation numérique vise à optimiser l'efficacité opérationnelle et à moderniser la gestion de ses ressources, y compris au sein de son département des Ressources Humaines.



Problématique

Perte de temps considérable

Le traitement manuel des CV, diplômes et autres documents administratifs est chronophage, détournant les équipes RH de tâches à plus forte valeur ajoutée.

Difficulté d'analyse et de reporting

L'absence de données structurées rend complexe l'extraction d'informations pertinentes pour l'analyse des effectifs, la planification des carrières ou la conformité réglementaire.

Risques d'erreurs de saisie

La transcription manuelle des informations est sujette aux erreurs humaines, pouvant entraîner des inexactitudes dans les dossiers du personnel.

Manque de standardisation des données

Les informations sont souvent stockées sous des formats hétérogènes, compliquant leur consolidation et leur exploitation uniforme.

Objectifs du stage



Objectif principal

Développer un système automatisé d'extraction de données structurées à partir de documents RH non structurés



Objectifs spécifiques



Classification automatique

Mettre en place une solution de classification automatique des documents RH (CV, diplômes, CIN) pour une meilleure organisation



Extraction de données

Implémenter un système d'extraction de données structurées et précises à partir des documents RH traités



Interface utilisateur

Développer une interface utilisateur intuitive et conviviale pour faciliter l'interaction avec le système



Stockage sécurisé

Assurer le stockage structuré et sécurisé des données extraites dans une base de données pour une exploitation ultérieure

Technologies utilisées



Langages

- Python 3.x



Librairies

- TensorLake SDK
- Pydantic
- pytesseract
- pdf2image
- pandas



Base de données

- SQLite



Interface

- Gradio



OCR

- Tesseract
- Support français/anglais



Stack Technologique



Intelligence Artificielle



Documents RH



Données Structurées

Méthodologie de travail

Approche: Développement agile avec itérations courtes

Une méthodologie itérative axée sur la livraison progressive de fonctionnalités fonctionnelles, avec une rétroaction continue et une adaptation flexible aux exigences changeantes.

Phases du projet



Analyse des besoins

Étude de l'existant et identification des besoins des utilisateurs



Conception

Architecture et modélisation des données



Implémentation

Développement des différents modules du système



Tests

Validation et tests unitaires/intégration



Documentation

Rédaction de la documentation technique et utilisateur

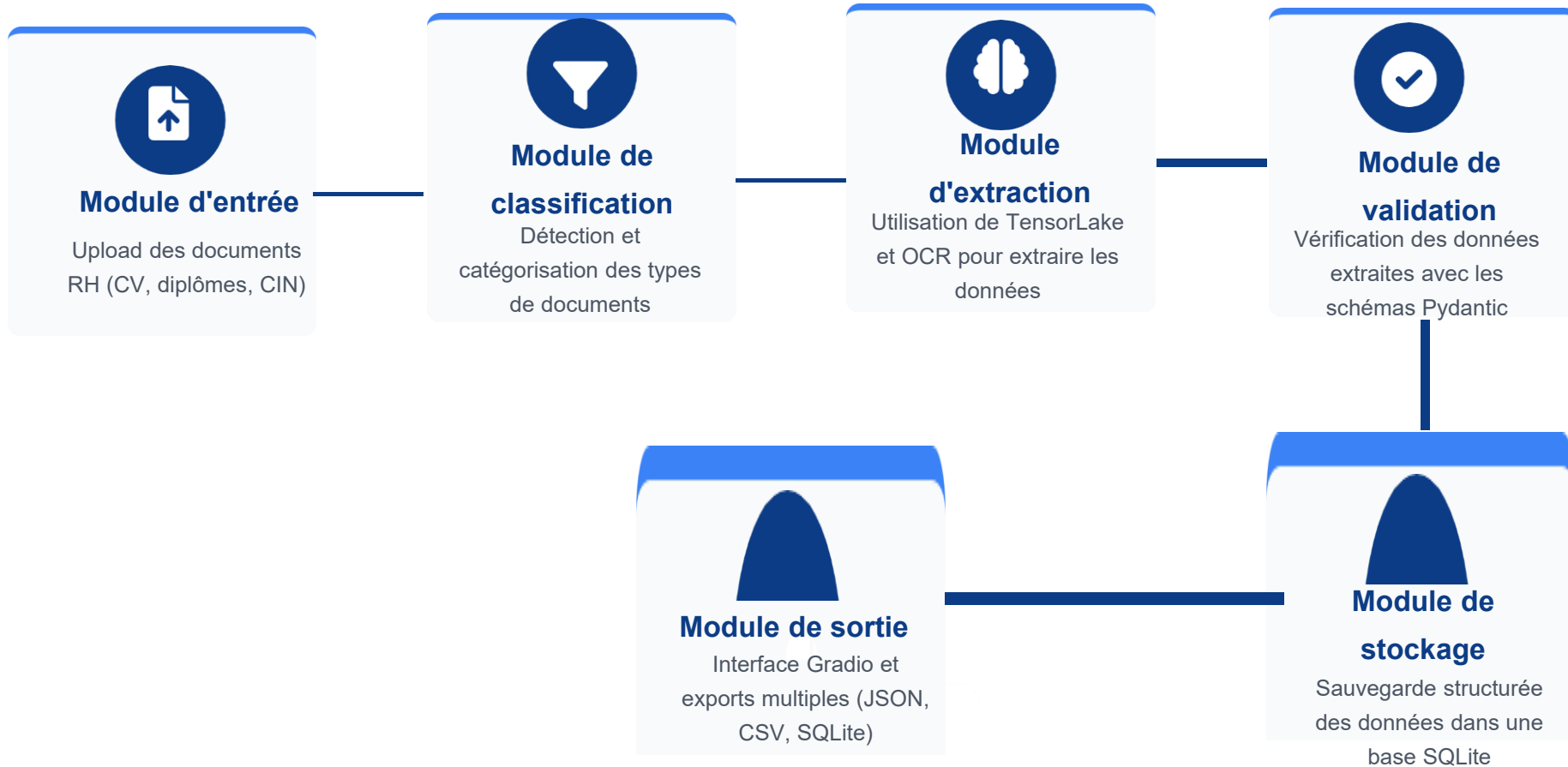
Avantages de l'approche agile



- Livraison progressive des fonctionnalités
- Rétroaction continue des utilisateurs
- Adaptation aux changements de spécifications
- Itérations courtes (1 mois)

Architecture du système

Flow Architecture du Système d'Extraction et de Structuration de Documents RH



Implémentation technique (1/2)

Modélisation des données avec Pydantic pour la structuration des documents RH



CVDData

Modèle pour les fichiers CV des candidats

Champ	Type
name	str
surname	str
email	str
phone	str
education	List[str]
skills	List[str]



DiplomaData

Modèle pour les diplômes

Champ	Type
degree	str
institution	str
student_name	str
graduation_year	int



CINData

Modèle pour les cartes d'identité nationales

Champ	Type
full_name	str
cin_number	str
birth_date	date
address	str
expiration_date	date

Pydantic a été choisi pour sa capacité à définir des schémas de données robustes et à valider les données en temps réel

Processus d'extraction des données des documents RH



1. Classification

Classification initiale des documents en analysant le nom du fichier et en réalisant une première passe OCR pour identifier des mots-clés spécifiques



2. Extraction

Utilisation du SDK TensorLake combiné à l'OCR (Tesseract) pour extraire les informations clés définies dans les schémas Pydantic



3. Validation

Validation des données brutes par rapport aux modèles Pydantic correspondants pour s'assurer qu'elles respectent les types et formats attendus



4. Stockage

Stockage des données validées de manière structurée dans une base de données SQLite locale, chaque type de document correspondant à une table spécifique



5. Export

Génération d'exports des données extraites dans divers formats (JSON, CSV, base SQLite complète) pour intégration avec d'autres systèmes

Avantages du processus d'extraction:

- Extraction fiable et validation rigoureuse des données
- Gestion structurée et sécurisée des informations
- Flexibilité pour l'intégration avec d'autres systèmes
- Capacité à traiter les CV, diplômes et cartes d'identité nationales

Interface utilisateur

Téléchargez un document RH pour extraire les données et récupérer automatiquement le CSV ou la base SQLite.

Uploader un document

HamzaElmourabitResume (2).pdf136.9 KB

Clear

Submit

Données extraites

```
1  ▼ [
2    ▼ {
3      ▼ "data": {
4        ▼ "education": [
5          ▼ {
6            "degree":
7              "2ème année préparatoire et Cycle d'ingénieur
              en Big Data et Système d'information"
8
9            "institution":
10             "Ecole Nationale des Sciences Appliquées de
11             Berrechid"
12
13            "year": "09/2023 - 06/2027"
14          },
15          ▼ {
16            "degree":
17              "1ère année du cycle préparatoire intégré",
18            "institution":
19              "Ecole Nationale des Sciences Appliquées de
20              Berrechid"
21          }
22        ]
23      }
24    }
25  ]
```

Interface utilisateur

Téléchargez un document RH pour extraire les données et récupérer automatiquement le CSV ou la base SQLite.

Uploader un document

HamzaElmourabitResume (2).pdf

136.9 KB ↓

Clear

Submit

Télécharger le CSV

extraction.csv

1.1 KB ↓

Télécharger la base SQLite

rh.db

8.0 KB ↓

Flag

Données extraites

```
1  ▼ [
2    ▼ {
3      ▼ "data": {
4        ▼ "education": [
5          ▼ {
6            "degree":
7              "2ème année préparatoire et Cycle d'ingénieur
8              en Big Data et Système d'information"
9            ,
10           "institution":
11             "Ecole Nationale des Sciences Appliquées de
12             Berrechid"
13           ,
14           "year": "09/2023 - 06/2027"
15         },
16       ▼ {
17         "degree":
18           "1ère année du cycle préparatoire intégré",
19         "institution":
20           "Ecole Nationale des Sciences Appliquées de
21           Berrechid"
```

Résultats obtenus

Performance du système d'extraction et de structuration des documents RH



Précision d'extraction

90%+

Taux de correspondance des données extraites
avec les informations réelles



Temps de traitement

30s-

Durée moyenne nécessaire pour traiter un
document complet



Taux de réussite global

85%

Proportion de documents traités avec succès sur
l'ensemble des tests



i Le système a été testé sur divers documents RH, notamment des CV, diplômes et cartes d'identité nationales. Les performances obtenues démontrent la robustesse du système pour l'automatisation des tâches d'extraction de données RH.

Difficultés rencontrées

Le développement de ce système a présenté plusieurs défis, tant sur le plan technique que méthodologique



Difficultés Techniques



Gestion des formats de documents variés

Documents RH se présentant sous une multitude de formats (scannés, numériques, différentes mises en page), nécessitant une adaptation constante des processus de pré-traitement et d'OCR



Optimisation de la qualité OCR

Précision de l'OCR cruciale pour les documents manuscrits ou de faible qualité, nécessitant des ajustements des paramètres de Tesseract et des étapes de nettoyage d'image



Validation des données extraites

Assurer l'exactitude des données extraites a été un défi majeur, avec la gestion des cas limites et des erreurs d'extraction nécessitant une attention particulière



Difficultés Méthodologiques



Intégration avec l'environnement existant de l'ONEE

Comprendre et s'adapter aux spécificités des processus RH et des systèmes d'information de l'ONEE a été un apprentissage continu, nécessitant une communication régulière avec les équipes internes



Respect des délais serrés (1 mois)

La courte durée du stage a imposé une gestion de projet rigoureuse, avec des itérations courtes et une priorisation stricte des fonctionnalités pour livrer un produit fonctionnel dans les temps impartis



Améliorations Techniques



Support de plus de types de documents

Étendre la capacité du système à traiter d'autres documents RH tels que les contrats de travail, les fiches de paie ou les attestations diverses



Intégration de modèles de langue (LLMs)

Exploiter la puissance des grands modèles de langage pour une extraction sémantique plus fine, une meilleure compréhension contextuelle et une gestion plus flexible des variations de mise en page



Amélioration de la précision OCR

Continuer à affiner les techniques d'OCR, potentiellement en explorant des moteurs OCR plus avancés ou des modèles d'apprentissage profond dédiés à la reconnaissance de texte



Déploiement



Intégration avec le SIRH de l'ONEE

Développer des connecteurs pour une intégration transparente du système avec le Système d'Information des Ressources Humaines (SIRH) existant de l'ONEE, permettant un flux de données automatisé



Mise en production avec interface web complète

Transformer l'interface Gradio en une application web robuste et sécurisée, offrant une expérience utilisateur enrichie et des fonctionnalités de gestion des documents plus avancées



Et bien d'autres possibilités d'expansion pour adapter le système à d'autres besoins organisationnels

Apports du stage



Pour l'ONEE



Gain de temps

Réduction considérable du temps passé par les équipes RH sur des tâches manuelles et répétitives



Réduction des erreurs

Standardisation et validation des données extraites minimisent les risques d'erreurs humaines



Amélioration de la qualité des données

Structuration des informations dès l'extraction, facilitant l'analyse et le reporting



Pour l'étudiant



Acquisition de compétences

Approfondissement des connaissances en intelligence artificielle, OCR et extraction d'informations structurées



Expérience en milieu professionnel

Immersion concrète dans les défis et exigences d'un grand organisme, application de compétences techniques à des problèmes réels



Développement de solutions end-to-end

Conception et mise en œuvre d'un système complet, de l'analyse des besoins à la démonstration d'une interface fonctionnelle

Conclusion

Ce projet de fin de stage a permis de développer un "Système Intelligent d'Extraction et de Structuration de Documents RH par IA" qui répond efficacement à la problématique de la gestion manuelle des documents à l'ONEE.

✓ **Objectifs atteints**

Nous avons atteint nos objectifs en mettant en place une solution automatisée capable de classer, extraire et structurer les données de CV, diplômes et CIN avec une précision et une efficacité notables.

💡 **Réponse à la problématique**

Le système met fin à la perte de temps dans le traitement manuel, à la saisie d'erreurs et à la difficulté d'analyse des données RH, en standardisant les informations pour une exploitation optimale.

📈 **Avancement significatif**

Ce système représente une avancée significative vers la digitalisation des processus RH, offrant un gain de temps considérable, une réduction des erreurs et une amélioration de la qualité des données.

🚀 **Voie ouverte**

Il ouvre la voie à une gestion des ressources humaines plus agile et plus stratégique au sein de l'ONEE, contribuant à la modernisation de ses processus métiers.

En conclusion, ce projet a permis de mettre en place une solution technique innovante et pratique qui répond aux besoins de l'ONEE en matière de gestion des documents RH, avec un impact direct sur l'efficacité opérationnelle du département des Ressources Humaines.

Veillez suivre la démonstration en direct du fonctionnement du système



Lien vers l'application Gradio : <https://2c59e97f072aefc4b6.gradio.live/>



Upload de documents

Démonstration du processus d'upload simple par drag-and-drop des documents RH



Extraction de données

Affichage des résultats d'extraction des données structurées en temps réel



Exports multiples

Démonstration des options d'exportation (JSON, CSV, base SQLite) des données extraites



À propos de la démonstration

Cette démonstration illustrera le processus d'upload, l'extraction des données et l'affichage des résultats sur des exemples concrets de documents RH (CV, diplômes, cartes d'identité nationales). Vous pourrez ainsi voir directement le fonctionnement du système et ses capacités d'analyse.

C: > Users > khadi > Downloads > Interface_Web_d'Extraction_de_Données_pour_Portail_RH (2).ipynb > M Installing dependencies

Generate + Code + Markdown Run All Clear All Outputs Outline ...

```
[1]
...
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  poppler-utils
0 upgraded, 1 newly installed, 0 to remove and 35 not upgraded.
Need to get 186 kB of archives.
After this operation, 697 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 poppler-utils amd64 22.02.0-2ubuntu0.10 [186 kB]
Fetched 186 kB in 0s (1,421 kB/s)
Selecting previously unselected package poppler-utils.
(Reading database ... 126435 files and directories currently installed.)
Preparing to unpack .../poppler-utils_22.02.0-2ubuntu0.10_amd64.deb ...
Unpacking poppler-utils (22.02.0-2ubuntu0.10) ...
Setting up poppler-utils (22.02.0-2ubuntu0.10) ...
Processing triggers for man-db (2.10.2-1) ...
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
tesseract-ocr is already the newest version (4.1.1-2.1build1).
0 upgraded, 0 newly installed, 0 to remove and 35 not upgraded.
Collecting pdf2image
  Downloading pdf2image-1.17.0-py3-none-any.whl.metadata (6.2 kB)
Collecting pytesseract
  Downloading pytesseract-0.3.13-py3-none-any.whl.metadata (11 kB)
...
Downloading pdf2image-1.17.0-py3-none-any.whl (11 kB)
Downloading pytesseract-0.3.13-py3-none-any.whl (14 kB)
Installing collected packages: pytesseract, pdf2image
Successfully installed pdf2image-1.17.0 pytesseract-0.3.13

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Merci de Votre Attention